文章编号:1671-4598(2017)07-0249-04 DOI:10.16526/j.cnki.11—4762/tp.2017.07.062 中图分类号:TP3 文献标识码:A

数据挖掘在肺结核疾病智能决策中的应用研究

王 科,马 丽,陈桂芬
(吉林农业大学 信息技术学院,吉林 长春 130118)

摘要:针对单一数据挖掘方法对肺结核疾病诊断效率低、准确性不高的问题,本研究对北京市昌平区结核病防治所,北京市结核病控制研究所的1203例肺结核病人档案资料构建了电子档案,采用粗糙集和决策树结合方法建立肺结核疾病诊断模型,并对其准确性进行评估,在此基础上构建肺结核疾病诊断系统;在研究中,使用粗糙集和决策树相结合的方法进行属性约简,约简掉冗余属性57个,剩余属性22个,得到决策规则7条,模型准确率为83.46%;与未未约简的方法相比,决策规则减少128%,模型准确率基本保持不变;研究结果表明:使用该组合算法,在保证模型准确率的同时,降低了算法的时间和空间复杂性,提高了挖掘效率,为临床诊断提供了一定的借鉴。

关键词:肺结核疾病;粗糙集;决策树;智能诊断

Application of Data Mining in Intelligent Decision of Pulmonary Tuberculosis Diseases

Wang Ke, Ma Li, Chen Guifen
(College of Information Technology, Jilin Agricultural University, Changchun 130118, China)

Abstract: Aiming at the problem that the low diagnostic efficiency and low accuracy of the single data mining method for Diagnosis of pulmonary tuberculosis, In this study, the electronic records of 1203 cases of tuberculosis patients in Changping District City, Beijing City of Beijng and Beijing Institute of tuberculosis control and tuberculosis control were build, Tuberculosis disease diagnosis model is built by application of rough set and decision tree method, On the basis of this, the diagnosis system of pulmonary tuberculosis was constructed. In this study, The combining method of rough set and decision tree was approached to attribute reduction, the model reduced redundant 57 attributes and remained 22 attributes, and articled 7 the decision rules. The model accuracy is 89.46%. Compared with the non reduction method, the decision rule was reduced by 128%, and the accuracy of the model remained unchanged. The research results showed that the algorithm can reduce the time and space complexity of the algorithm while ensuring the accuracy of the model, so as to improve the efficiency of the mining, and provide some references for clinical diagnosis.

Keywords: pulmonary tuberculosis disease; rough set; decision tree; intelligent diagnosis

0 引言

医疗设备和仪器的数字化,使得医院数据库的信息容量不断地膨胀,包括大量关于病人的病史、诊断、检验和治疗的临床信息。如何通过高效、智能的计算机算法对海量肺结核疾病诊疗数据进行数据挖掘,根据治疗结果与医疗过程中的病历数据之间的隐藏关系,寻找可行可靠的诊疗方法,及早有效的为医务人员提供针对性的辅助诊疗方案,具有重要的临床意义。

肺结核是严重危害人类健康的慢性呼吸道疾病。目前典型肺结核的诊断主要病历数据之间的隐藏关系,寻找可行可靠的诊疗方法,及早有效的为医务人员提供针对性的辅助诊疗方案,具有重要的临床意义。

肺结核是严重危害人类健康的慢性呼吸道疾病。目前典型肺结核的诊断主要通过对临床表现的观察、痰查结核菌、胸部影像学、支气管镜检查等做出判断。如何对就诊的肺结核可疑症状者及疑似肺结核患者进行合理检查及早期诊治,减少结核菌的进一步传播;如何摆脱单一指标,建立患者多模态临床信息,并从中挖掘出与病理密切相关的临床指标,实现对肺结核疾病的临床鉴别,是呼吸科重要的临床需求之一。

本文尝试根据肺结核数据特点改进现有挖掘算法,运用粗糙集和决策树方法相结合的方法,进行属性约简,并提取决策树规则集,挖掘病历数据用隐含的诊断规则,获取新的知识发现,为结核病患者的临床治疗提供参考。

收稿日期:2017-02-06; **修回日期**:2017-04-12。
基金项目:国家星火计划(2015GA66004)。
作者简介:王 科(1985-),男,北京人,硕士研究生,主要从事人工智能与计算机应用方向的研究。
通讯作者:陈桂芬(1956-),女,教授,博士,博士生导师,主要从事人工智能与数据挖掘,精准农业方向的研究。

1 数据挖掘基础

1.1 粗糙集理论

粗糙集以知识、近似集合等概念为核心处理不精确、不确定与不完全数据的理论。在问题处理中,不需要提供问题数据集以外的任何先验信息,因此能较客观的处理不确定性问题。

粗糙集理论中采用四元有序组描述知识,即:$K=(U, A, V, d)$。其中U是论域;A是属性全体;$V=\bigcup a\in A\ Va$,Va是属性的值域;$d: U\times A\rightarrow V$是一个信息函数,$dx: A\rightarrow V$,$x\in U$,表示了对象$x$在$K$中的完全信息,其中$dx(a)=d(x, a)$。对于这样的信息系统,每个属性子集就定义了论域上的一个等价关系,即BA,定义$R_B: xR_B\ y\ dx(b)=dy(b)$,$b\in B$。

记由属性集$B\subseteq A\ T$所导出的等价关系为R_B。$a\in A\ T$,如果$R_{AT}=R_{AT\setminus\{a\}}$,则称属性$a$是多余的;如果在系统中没有多余属性,则称$AT$是独立的;如果$R_B=R_{AT}$且$B$中没有

多余属性子集，则 B_{AT} 称为是 $A\ T$ 的约简，记作 $red\ (\ AT\ )$；$A\ T$ 的所有约简的交集称为 $A\ T$ 的核，记作 $core\ (\ AT\ )$。

1.2 决策树方法

决策树方法通过对大量数据按一定目标进行分类，将从一组训练数据中学习到的函数表示为一棵决策树，从中找到有用的、潜在的信息，常用于分类预测的算法。决策树方法具有速度快、精度高、生成的模式简单等特点，在数据挖掘具有广泛的应用。

构造决策树是包括两个步骤：生成决策树和决策树剪枝。生成决策树时是从一个根节点开始，通过不断地将样本分割成子集，进行从上到下的递归过程构造出一棵树。对每个属性的测试取值表示为树上的非叶结点，每个结果表示为树的一个分枝，最终的分类类别为树的叶子结点。决策树构造中，使用信息增益作为对节点进行划分的标准。

由于有噪声数据和孤立点，因此生成的决策树会引起分枝异常，故需要对决策树进行剪枝。在决策树剪枝中，通常选用叶结点来代替一个或多个子树，然后选择概率最高的类为该结点的类别，也可以用其中的树枝来代替子树。

2 模型构建

2.1 电子病历构建

原始样本数据来源于北京市昌平区结核病防治所，北京市结核病控制研究所病历档案，数据采集时间为 2015 年 11 月～2016 年 5 月，应用 Microsof t SQL2010 对来自不同数据源的数据进行整合，涉及病历档案资料 1203 份。与本项研究有关数据主要三大类：

（1）患者一般信息：①病历号、②性别、③出生日期、④分组（初治组、复治组）、⑤户籍类型（本市、外省）、⑥民族（汉族、回族、满族、其他）、⑦密接史（无、有）、⑧既往有无合并其他疾病（糖尿病、矽肺、肝炎、癫痫、肺癌、肺部感染、肺心病、慢支、其他）⑨既往有无肺外结合病史（结核性胸膜炎、淋巴结核、骨结核、皮肤结核、肾结核、腹膜结核、盆腔结核、肠结核、输卵管结核）

（2）疗前主要症状：②咳嗽、咳痰≤2 周②①咳嗽、咳痰＞2 周③咯血/痰中带血④胸痛⑤午后低热⑥盗汗⑦乏力⑧、食欲减退⑨体重减轻▮月经不调▮体检发现，无任何症状▮其他

（3）疗前检查项目开展情况：①血沉②C 反应蛋白（无，有）③疗前痰抗酸杆菌涂片（未查、已做结果）④疗前痰抗酸杆菌普通培养（未查、已做结果）⑤痰结核分枝杆菌快速培养（未查、已做结果）⑥培养阳性患者菌种鉴定（传统/快速）（未查、已做结果）⑦结核菌素试验（未查、已做结果）⑧结核抗体（未查、已做结果）⑨γ－干扰素释放试验/T－SPOT（未查、已做结果）⑩痰结核杆菌定量 PCR（未查、已做结果）▮痰结核杆菌 Hain 试验（未查、已做结果）▮痰结核杆菌 X－pert 检测（未查、已做结果）▮血液肿瘤标志物检查（未查、已做结果）▮支气管镜检查（未查、已做结果）▮活检（肺组织/胸膜/胸水）（未查、已做结果）▮疗前胸部 DR（未查、已做结果）▮疗前胸部 CT（未查、已做结果）（4）最终诊断（肺结核/胸膜炎、不是肺结核、NTM）。原始数据如图 1 所示。

2.2 基于数据挖掘的结核病诊断

2.2.1 数据预处理

由于病历是由医生或非医学专业人员手工录入文本或数据库，会存在数量大、记录形式不统一、记录错误和噪声，因此需要对原始病历数据进行数据预处理。数据预处理主要包括病例数据采集，属性选择，连续属性离散化，数据中的噪声及丢失值处理，实例选择等。为进一步进行数据挖掘，需要对信息数据表中的值和字段进行编码，对于肺结核疾病的编码如表 1 所示，预处理后的数据如图 2 所示。

来源	病历号	性别	出生日期	分组	户籍类型	民族	密接史	既往有无合	咳嗽、咳痰≤2周	>2周	血沉值	阳性，数值	阳性，数值	最终诊断
市所	84122	男	05/11/1999	初治组	外省	汉族	无	无	有	有	无	2	16	216	肺结核/胸膜炎
市所	84121	女	01/12/1986	初治组	外省	汉族	无	无	无	无	无				肺结核/胸膜炎
市所	84120	男	10/03/1987	初治组	本市	汉族	无	无	无	无	无	5	88	4	肺结核/胸膜炎
市所	84119	男	12/05/1990	初治组	外省	汉族	无	无	有	无	有	61			肺结核/胸膜炎
市所	84117	女	01/02/1994	初治组	外省	汉族	无	无	无	无	有		24	24	肺结核/胸膜炎
市所	84115	女	04/11/1942	初治组	本市	汉族	无	无	无	无	无	5			肺结核/胸膜炎
市所	84114	女	01/07/1949	初治组	外省	汉族	无	有	无	无	无	44			肺结核/胸膜炎
市所	84113	女	05/06/1939	初治组	本市	回族	无	无	无	无	无	9	408	288	不是肺结核
市所	84176	女	06/01/1986	初治组	外省	汉族	无	无	无	无	无	14	140	232	肺结核/胸膜炎
市所	84175	女	20/11/1991	初治组	外省	汉族	有	无	无	无	无	2			不是肺结核

图 1 部分原始数据

表 1 肺结核疾病属性编码表

属性	数据离散化及编码	属性	数据离散化及编码
性别	(1)男 (2)女	咳嗽、咳痰	(1)无 (2)有
分组	(1)初治组 (2)复治组	咳嗽、咳痰≤2 周	(1)无 (2)有
户籍类型	(1)本市 (2)外省	盗汗	(1)无 (2)有
民族	(1)汉族 (2)回族 (3)满族 (4)其他	午后低热	(1)无 (2)有
密接史	(1)无 (2)有	咯血/痰中带血	(1)无 (2)有
既往有无合并其他疾病	(1)无 (2)有	疗前痰抗酸杆菌涂片	(1)未查 (2)已做结果
糖尿病	(1)无 (2)有	疗前痰抗酸杆菌普通培养	(1)未查 (2)已做结果

来源	病历号	性别	出生日期	分组	户籍类型	民族	密接史	既往有无合	咳嗽、咳痰≤2周	>2周	血沉值	阳性，数值	阳性，数值	最终诊断
市所	84171	2	15/02/1990	1	2	1	1	1	1	undifinec	undifinec	45	780	2520	1
市所	84170	1	25/04/1990	1	1	1	2	1	2	undifinec	2	25	56	172	1
市所	84169	2	02/10/1993	1	2	1	1	1	1	undifinec	undifinec	21	128	152	1
市所	84168	1	22/10/1976	1	2	1	1	1	1	undifinec	undifinec	23	undifinec	undifinec	1
市所	84169	1	07/09/1977	1	2	1	1	1	1	undifinec	undifinec	56	116	120	1

图 2 预处理后部分数据

2.2.2　基于粗糙集的属性约简

在病历数据中属性较多，而各个属性之间往往存在着某种程度上的依赖关系，不能简单的删除。约简在不丢失信息的前提下，能较简单地表示决策系统的决策属性集合对条件属性集合的依赖关系，能够从条件属性中去掉不必要的条件属性，简化条件属性，提高挖掘效率。本文中，对 76 个属性进行约简，约简后剩余属性 22 个，约简掉冗余属性 57 个。约简后的属性如表 2 所示。

表 2　约简后属性

患者一般信息(3 项)	密接史、糖尿病、结核性胸膜炎
疗前主要症状(5 项)	咳嗽、咳痰≤2 周、咯血/痰中带血、胸痛、体检发现，无任何症状、其他
疗前检查项目开展情况(14 项)	血沉值、C 反应蛋白值、疗前痰抗酸杆菌普通培养已做结果、痰结核分枝杆菌快速培养已做结果、结核菌素试验、结核抗体已做结果、γ－干扰素释放试验/T－SPOT、痰结核杆菌 Hain 试验、痰结核杆菌 X－pert 检测、血液肿瘤标志物检查、支气管镜检查、活检(肺组织/胸膜/胸水)、疗前胸部 DR 已做结果、疗前胸部 CT 已做结果

2.2.3　决策树模型建立

按照信息增益建立属性重要度，如图 3 所示。按从大到小顺序依次为：疗前痰抗酸杆菌普通培养已做结果、痰结核分枝杆菌快速培养结果、疗前胸部 DR 已做结果、结核抗体结果、痰结核杆菌 X－pert 检测、结核菌素试验、γ－干扰素释放试验/T－SPOT、活检（肺组织/胸膜/胸水），这与临床上确诊肺结核患者诊疗标准基本一致。

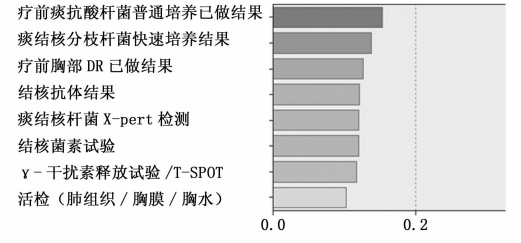

图 3　属性重要度排序

按 C5.0 构建决策树，为避免单次分区的抽样误差对结果的影响，提高模型准确率，在决策树建模的时进行十折交叉验证，模型正确率 83.46%，支持数据条数 1，004。挖掘出的决策规则 7 条，按照支持度和置信度降序排序如表 3 所示。

2.3　结果分析

从挖掘出的规则中，大部分样本都集中于置信度 60%～99% 的数值内，在从实际临床诊断结果中，实例数和置信度比较高即强关联时这一区间包含的规则数据，对肺结核疾病诊断具有较高的临床价值。本研究模型已在北京市昌平区结核病防治所，北京市结核病控制研究所进行应用，效果良好。

本研究对粗糙集与决策树相结合的方法与单一决策树方法，在规则数目、准确率、置信度区间和建模时间四个方面进行了对比，对比结果如表 4 所示。从表中可以看出单一的决策树方法在分类预测中，还存在一定冗余属性，致使构造出的决策树规模较大，提取的规则较多，导致决策时挖掘效率不高。本研究利用已有的肺结核疾病档案数据，利用粗糙集与决策树相结合的优化算法对筛选后的 22 个属性变量建立结核病治疗的预测与分类模型，去掉了冗余属性，简化了决策模型，提高了挖掘效率。

表 3　按照支持度和置信度降序排序的规则结果

规则	实例数	置信度%
如果 疗前痰抗酸杆菌普通培养已做结果＞ 1 则 肺结核/胸膜炎	269	98.2
如果疗前胸部 DR 已做结果 ＜＝ 2 则不是肺结核	234	96.7
如果 结核抗体已做结果 ＞ 1 并且 疗前胸部 DR 已做结果＞ 2 则 肺结核/胸膜炎	32	94.1
如果结核菌素试验＜＝ 1 并且结核抗体已做结果＜＝ 1 并且 γ－干扰素释放试验/T－SPOT ＜＝ 1 并且 痰结核杆菌 X－pert 检测 ＞ 1 并且 疗前胸部 DR 已做结果 ＞ 2 则 肺结核/胸膜炎	1 050	85.3
如果疗前胸部 DR 已做结果 ＞ 2 则 肺结核/胸膜炎	121	74.1
如果 疗前痰抗酸杆菌普通培养已做结果＜＝ 1 并且 痰结核分枝杆菌快速培养已做结果 ＜＝ 1 并且结核菌素试验＜＝ 1 并且 γ－干扰素释放试验/T－SPOT ＜＝ 1 并且 活检（肺组织/胸膜/胸水）＜＝ 1 则 不是肺结核	338	71.5
如果 疗前痰抗酸杆菌普通培养已做结果 ＜＝ 1 并且结核菌素试验＜＝ 1 并且 γ－干扰素释放试验/T－SPOT ＜＝ 1 则 不是肺结核	593	60.7

表 4　模型对比结果

模型	规则数目	准确率	置信度区间	建模时间
组合优化方法	7	83.46%	60%～99%	0.02
单一决策树方法	16	82.18%	63%～98%	0.25

2.4　智能诊断系统构建

本文构建的肺结核疾病诊断系统，在对数据进行预处理后，应用基于决策树与粗糙集相结合的方法构建决策模型。在应用中，输入患者信息，应用该决策模型进行智能决策，得到疾病诊断结果。

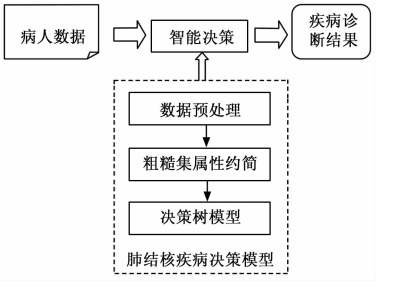

图 4　肺结核疾病诊断系统体系结构

3　结论

数据挖掘中决策树方法在肺结核疾病诊断中已有应用，如张琪的“决策树模型用于结核病治疗方案的分类和预判”，说明该数据挖掘方法适用于肺结核疾病分类诊断问题，但单一应用决策树构建诊断模型，纳入研究的变量数较多，可能会引起检验功效降低的问题。

计算机辅助医学数据挖掘实现了医学数据的冗余性消除、

规范化储存以及数据无缝融合与共享、知识提取的自动化与可视化表达等功能。本研究利用已有的肺结核疾病档案数据，利用粗糙集与决策树相结合的优化算法对筛选后的 22 个属性变量建立结核病治疗的预测与分类模型，去掉了冗余属性，提高了挖掘效率，为临床诊断提供一定的借鉴。此外，本研究的研究思路和方法也可以适用于其他的慢性疾病治疗方案的选择，例如高血压、糖尿病。

医学的发展已经由经验医学、实验医学转向以证据为基础的循证医学，产生的医学数据量大，且具有客观性。临床医生应结合患者的病史、临床症状，逐渐学会使用大样本下建立的数据挖掘模型，分析病情、制定合理治疗方案及预测疾病的发展。实践证明，基于粗糙集和决策树相结合的优化算法，可以有效地处理数据中的不确定性推理，是数据挖掘的有力工具，在今后的医疗数据资源挖掘利用中具有广阔的应用前景。

参考文献：

[1] 王　欣，翁少凡，祖爱华，等. 肺结核危险因素病例对照研究 [J]. 工业卫生与职业病，2011 (4)：208－213.

[2] 陈桂芬，马　丽，董　玮，等. 聚类、粗糙集与决策树的组合算法在地力评价中的应用 [J]. 中国农业科学，2011，23：4833－4840.

[3] 石　凯. 基于粗糙集理论的属性约简与决策树分类算法研究 [D]. 大连：大连海事大学，2014.

[4] 王国胤，姚一豫，于　洪. 粗糙集理论与应用研究综述 [J]. 计算机学报，2009 (7)：1229－1246.

[5] 张　明. 粗糙集理论中的知识获取与约简方法的研究 [D]. 南京：南京理工大学，2012.

[6] 冯少荣. 决策树算法的研究与改进 [J]. 厦门大学学报 (自然科学版)，2007 (4)：496－500.

[7] 卢东标. 基于决策树的数据挖掘算法研究与应用 [D]. 武汉：武汉理工大学，2008.

[8] 王　俊，马　丽，赵　叙，等. 数据挖掘技术在建立肺结核病单病种费用模型中的研究 [J]. 中国医院，2014 (4)：33－34.

[9] 张宏宇. 粗糙集理论及其在数据挖掘中的应用研究 [D]. 太原：山西大学，2003.

[10] 陆维嘉. 辅助慢性呼吸道疾病诊疗的电子病历系统与数据挖掘研究 [D]. 上海：上海大学，2015.

[11] 高军舰. 关联规则在呼吸内科诊疗数据中的应用研究 [D]. 青岛：青岛科技大学，2014.

[12] 张　琪，周　琳，陈　亮，等. 决策树模型用于结核病治疗方案的分类和预判 [J]. 中华疾病控制杂志，2015 (5)：510－513.

[13] 任正洪. 2005～2011 年我国肺结核发病的时间流行病学特征及趋势 [J]. 中国卫生统计，2013 (2)：158－161.

(上接第 248 页)

图 6　申请用电页面

若宿管人员需要查看本单位各个寝室房间的用电情况，在登录系统平台中输入自己的职工编号和密码后再输入本单位的编号，如 5 号楼，其系统反馈页面如图 7 所示，第一列为本单位中各个房间号，然后依次为每个房间每月可用电量，即学校每个期望学生用电 30 kW，如学生有超出部分，则需要学生自己支付超出部分的用电费用，当月已使用的电量，当前可申请短时间内可使用的限定功率最大值，即每次学生短时间内可使用的最大用电功率为 1.8 kW。当月的剩余电量，以及当日已申请的用电次数。每个寝室的每日可申请的用电次数是由系统管理员来设置的，同时单词允许使用大功率用电的时长也是由系统管理员来设置，这样就避免了学生们无限次连续的使用的使用大功率用电器，这样处理的目的是既方便了学生的用电，又可以有效的安全和节约用电。

4　结束语

远程用电管理系统是基于 B/S 架构开发的嵌入式系统的应用，它是依据当今社会各高校还在依赖人工管理的用电管理系统中存在的缺陷和弊端下提出的一套智能用电管理平台，将校园内的系统用电合理化重新分配，考虑到对于不同用户的安全、节约用电，远程用电系统对各个部门的用电申请提出了系统限制，保证各个部门合理的申请和查看本部门的用电情况的

图 7　管理员查询页面

基础上，同时系统终端还会对终端的工作环境进行检测，将检测到的数据信息进行系统上传，以及工作预警。远程用电管理系统将使所有用户同一平台分权限管理，降低同一平台多用户工作的耦合性，远程用电管理系统在保证了校园安全节约用电的基础上，智能化的对校园内各用电单位从新分配电能，对于各个用电单位合理的用单查看和申请不需要繁琐的审批流程，既提高了办事的快速性，也是工作变得灵活可靠。通过系统测试，远程用电管理系统在达到校园内安全节约用电的同时，还极大的方便了各用电部门灵活用电调度的工作，对于开放用电部门也改善了传统的用电浪费的现象。

参考文献：

[1] 王云亮，李　莹等. 基于嵌入式安防远程监控系统的设计 [J]. 化工自动化及仪表，2016，(42)：32－34.

[2] 陈　娟，李　元，李万国. 基于 B/S 架构的嵌入式系统测试方法与实现 [J]. 电子技术应用，2016，42 (2)：50－52，56

[3] 李志宇. 物联网技术研究进展 [J]. 计算机测量与控制，2012，6 (48)：65－67

[4] 蔡　勇，王　勇. 嵌入式数据库 SQLite 在测控系统设计中的应用 [J]. 电子测试，2011 (10)：89－93.

[5] 孙街亭，徐蓓蓓. B/S 架构的宿舍管理信息系统研究与开发 [J]. 西安工程大学学报，2010，24 (2)：197－202.