关联规则算法在临床医疗诊断中的应用

姚旭升[1],杨 静[1],谢颖夫[2],贺建峰[1]

(1.昆明理工大学 信息工程与自动化学院;2.云南省第一人民医院,云南 昆明 650000)

摘 要:为了从临床数据中挖掘出疾病之间的相关性,为疾病临床诊断提供一种辅助方法,使用SPSS Statistics进行数据预处理,将数据转化为布尔数据,最后应用SPSS Modeler搭建基于Apriori算法的关联规则挖掘数据流,采用云南某医院2013年住院病案首页数据(共54 841条)建立疾病间的关联规则模型。从227种疾病中挖掘出信度大于20%的关联规则共40条,涉及20种疾病。关联规则挖掘可以从大量临床数据中发现疾病间潜在关联,为相关疾病的临床诊断提供辅助。

关键词:数据挖掘;关联规则挖掘;SPSS Modeler;临床辅助诊断

DOI:10.11907/rjdk.172521

中图分类号:TP319 **文献标识码**:A **文章编号**:1672-7800(2018)003-0162-03

Application of Association Rule Algorithm in Clinical Medical Diagnosis

YAO Xu-sheng[1], YANG Jing[1], XIE Ying-fu[2], HE Jian-feng[1]

(1. *School of information engineering and automation, Kunming University of Science and Technology*;
2. *The first people's Hospital of Yunnan Province, Kunming* 650000, *China*)

Abstract: In order to dig out the correlation between diseases from clinical data, an auxiliary method is provided for the clinical diagnosis of disease. SPSS Statistics is used to preprocess the data and convert the data into Boolean data. Finally, SPSS Modeler is applied to build association rules mining data stream based on Apriori algorithm. An association rule model between diseases was established by using the first page of inpatient medical records(a total of 54 841) in a hospital in Yunnan in 2013. A total of 40 association rules with confidence setting greater than 20% were extracted from the 227 diseases, involving 20 diseases. Association rules mining can discover the latent association between diseases from a large amount of clinical data. This can provide an auxiliary method for the clinical diagnosis of related diseases.

Key Words: data mining; association rules; SPSS Modeler; clinical assistant diagnosis

0 引言

随着信息技术的发展,目前很多领域已经逐渐积累起海量数据,数据挖掘手段可以从这些数据中挖掘出一些人类不容易发现的潜在规律。数据挖掘可以概括为一种决策支持过程,主要基于人工智能、机器学习、统计学等技术,高度自动化地分析原有数据,作出归纳性推理,从中挖掘出潜在规律,预测分析对象的行为趋势,从而帮助决策或调整策略[1]。

关联规则算法是用来探索事务之间依赖关系的一种常用方法,最典型的应用是挖掘超市交易数据中售出商品间潜在关系,用于找出顾客购买行为模式,从而优化商品布置,以达到增长销售额的目的[2]。目前关联规则挖掘已广泛应用于各个行业。在医学领域,关联规则广泛应用于临床用药规律、疾病预测分析等方面[3]。关联规则算法的特点是可以发现自然组合的关联,将该方法应用于挖掘不同种疾病之间的相关性,对于疾病的主动预防以及临床辅助诊断是有意义的[4]。

本文基于SPSS Modeler软件提出一种针对住院病案首页中诊断数据的疾病相关性挖掘方法,采用云南省昆明市某三甲医院2013年住院病案首页数据,力图挖掘出一些

收稿日期:2017-09-12

基金项目:国家自然科学基金项目(11265007)

作者简介:姚旭升(1992-),男,昆明理工大学信息工程与自动化学院硕士研究生,研究方向为计算机应用;杨静(1993-),女,昆明理工大学信息工程与自动化学院硕士研究生,研究方向为医院信息化建设;谢颖夫(1965-),男,云南省第一人民医院信息中心主任,研究方向为医疗卫生系统计算机网络与医学数据分析;贺建峰(1965-),男,博士,昆明理工大学信息工程与自动化学院教授,研究方向为计算机应用、医院信息化建设等。本文通讯作者:贺建峰。

疾病间可能的潜在关联，为临床诊断提供帮助。

1　关联规则挖掘

1.1　关联规则挖掘定义

关联规则挖掘可描述如下：

设 $I=\{i_1,i_2,\ldots,i_m\}$ 是有 m 个不同的项组成的集合，简称项集。给定一个事务集合 D，其中每一个事务 T 是 I 中一组项的集合，即 $T\subseteq I$。若项集 $A\subseteq I$ 且 $A\subseteq T$，则事务 T 包含项集 A[5]。关联规则是形如 $A\rightarrow B$ 的关系式，其中 $A\cup T$，$B\cup T$，且 $A\cap B=\varnothing$；关联规则挖掘是要在事务集合 D 中找出所有满足最小支持度和最小置信度的关联规则。

1.2　Apriori 算法

Apriori 算法是一种最有影响的布尔关联规则频繁项集挖掘算法[6]，其核心是基于两阶段频集思想的递推算法。该关联规则在分类上属于单维、单层、布尔关联规则[7]，所有支持度大于最小支持度的项集称为频繁项集，简称频集。

该算法的基本思想：①找出所有频集，这些项集出现的频繁性至少与预定义的最小支持度一样；②由频集产生强关联规则，这些规则必须满足最小支持度与最小可信度；③使用第 1 步找到频集产生期望的规则，产生只包含集合项的所有规则，其中每一条规则右部只有一项。一旦这些规则被生成，那么只有那些大于用户给定的最小可信度的规则才被留下来。

3　关联规则挖掘方法

3.1　研究对象

本文采用的数据是云南省某三甲医院 2013 年全年的住院病案首页数据，共 54 841 条，根据住院病案首页国家标准，每个住院案例包含四大类指标，分别为患者基本信息、住院过程信息、诊疗信息与费用信息[8]。由于本文研究的是疾病间潜在的联系，故选择以下字段作为研究指标：住院病案号（为保护患者隐私，采用住院病案号作为患者身份标识）、疾病编码（主要诊断编码）、疾病编码 1（其它诊断 1 编码）……疾病编码 16（其它诊断 16 编码），共 17 个指标。

3.2　数据预处理

2013 年住院病案首页数据中把本文不考察的其它指标过滤掉，仅留下研究对象。采用 SPSS Statistics 22 将数据转化为事务处理格式，统计疾病频数，疾病频数小的疾病对于模型的影响微乎其微，故将疾病频数小于 100 的案例删除，提高建模效率。

3.3　关联规则挖掘模型构建

SPSS Modeler 软件的特点是采用数据流形式处理数据，可以直观地分析数据处理过程、设置参数[9-10]。本文采用 SPSS Modeler 14.1 建立关联规则挖掘数据流模型。模型如图 1 所示。

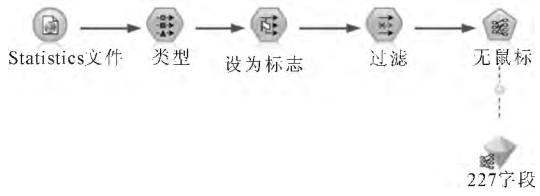

图 1　关联规则挖掘数据流模型

其中，在“Statistic 文件”节点中输入预处理后的数据

表 1　关联规则挖掘结果

后项	前项	支持度 %	置信度 %
高血压	多发性脑梗死	5.5	66.1
高血压	动脉粥样硬化	2.6	62.7
高血压	冠状动脉粥样硬化性心脏病	6.3	62.0
高血压	糖尿病	11.4	58.8
多发性脑梗死	动脉粥样硬化	2.6	54.4
慢性支气管炎	肺气肿	2.0	50.3
高血压	慢性支气管炎	2.2	49.7
高血压	慢性肾功能衰竭	2.6	49.0
高血压	前列腺增生	2.3	48.1
肺气肿	慢性支气管炎	2.2	46.7
高血压	高脂血症	2.4	46.2
高血压	慢性阻塞性肺疾病	3.3	44.1
冠状动脉粥样硬化性心脏病	慢性阻塞性肺疾病	3.3	43.9
慢性肾功能衰竭	继发性贫血	2.4	40.7
高血压	肺气肿	2.0	40.6
继发性贫血	慢性肾功能衰竭	2.6	36.9
糖尿病	高脂血症	2.4	35.2
胃炎	慢性胆囊炎	2.1	31.4
高血压	肺部感染	4.0	31.2
高血压	并发性白内障	2.1	29.8
糖尿病	多发性脑梗死 and 高血压	3.6	29.5
糖尿病	高血压	22.7	29.4
糖尿病	冠状动脉粥样硬化性心脏病 and 高血压	3.9	29.1
慢性胆囊炎	胃炎	2.3	28.5
冠状动脉粥样硬化性心脏病	肺气肿	2.0	27.8
冠状动脉粥样硬化性心脏病	前列腺增生	2.3	27.4
急性咽炎	新生儿高胆红素血症	2.1	26.7
动脉粥样硬化	多发性脑梗死 and 高血压	3.6	26.5
动脉粥样硬化	多发性脑梗死	5.5	25.6
糖尿病	多发性脑梗死	5.5	24.0
高血压	继发性贫血	2.4	23.7
糖尿病	冠状动脉粥样硬化性心脏病	6.3	22.8
慢性阻塞性肺疾病	冠状动脉粥样硬化性心脏病	6.3	22.8
糖尿病	动脉粥样硬化	2.6	22.4
新生儿红斑	新生儿高胆红素血症	2.1	22.0
肺部感染	慢性阻塞性肺疾病	3.3	21.9
冠状动脉粥样硬化性心脏病	慢性支气管炎	2.2	21.4
糖尿病	慢性肾功能衰竭	2.6	21.1
早产儿	新生儿高胆红素血症	2.1	20.9

源；在“类型”节点中将纳入模型的变量类型设为“输入”；在“标志”节点中将“疾病代码”设为标志字段。按照“住院病案号”进行汇总，其目的是将现有数据变成可以被布尔数据 Apriori 算法处理的数据。在“过滤”节点中将疾病编码修改为疾病中文名，最后在“Apriori 算法”节点中设置最小置信度为 20%，最小支持度为 2%，运行模型、输出关联规则结果 40 条。

4 结果

关联规则算法对疾病间相关性挖掘结果见表 1。从 227 种疾病中挖掘出 40 条关联规则。其中前项和后项的含义是若事件 A 存在的同时事件 B 存在，那么前项就是 A，后项是 B；支持度的含义是事件 AB 同时发生的实例占总案例的比例；置信度的含义是 AB 事件同时发生占事件 A 的比例。如第一条的意义是多发性脑梗死的患者同时患有高血压的占总案例数的 5.5%，多发性脑梗死患者中 66.1%的人同时患有高血压。其关联的内在原因有待医学专家进一步研究。

5 结语

数据挖掘在医学领域的应用前景十分广阔，本文应用 SPSS Modeler 软件，通过对某医院 2013 年住院病案首页数据的疾病相关性进行挖掘，给出了一个可行的关联规则挖掘实施方案，挖掘出一些可能有价值的关联规则。当数据量增大时，可能会从中挖掘出更多有价值的潜在联系。以上挖掘出部分关联可以为临床诊断提供辅助参考，同时对于疾病预防、宣传也有一定的积极作用。

参考文献：

[1] 应振潭. 数据挖掘技术在生源质量分析中的应用[J]. 软件导刊，2009(8)：172-173.

[2] 林犷. 慢性肾小球肾炎的中医症状-证候-药物关联规则挖掘的研究[D]. 成都：电子科技大学，2016.

[3] 赵佳璐. 基于关联规则挖掘的出生缺陷预警系统的研究与实现[D]. 北京：北京邮电大学，2012.

[4] 郑传生，蔡伟鸿. 一种关联规则挖掘算法及其在医疗信息挖掘中的应用[J]. 计算机与现代化，2007(7)：10-12.

[5] SAHOO J，DAS A K，GOSWAMI A. An efficient approach for mining association rules from high utility itemsets[J]. Expert Systems With Applications，2015，42(13)：5754-5778.

[6] AGRAWAL R，SRIKANT R. Mining sequential patterns[C]. IEEE Computer Society，1995：3-14.

[7] P TANNA，Y GHODASARA. Using apriori with WEKA for frequent pattern mining[J]. International Journal of Engineering Trends and Technology，2014，12(3)：127-131.

[8] 国家卫生计生委办公厅. 住院病案首页数据填写质量规范（暂行）[R]. 北京：2016

[9] 张文彤，钟云飞. IBM SPSS 数据分析与挖掘实战案例精粹[D]. 北京：清华大学出版社，2013.

[10] 季聪华，曹毅，张颖，等. 基于 SPSS Clementine 软件的关联规则算法的应用[J]. 中医药管理杂志，2014(1)：31-33.

（责任编辑：刘亭亭）

（上接第 161 页）

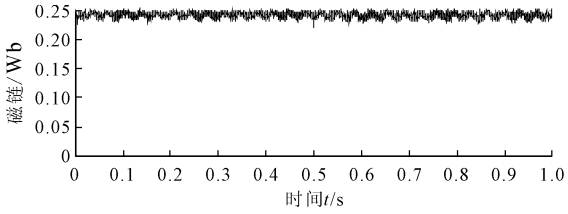

图 9 传统直接推力控制磁链波形

输出波形可以看出，系统引入 SVPWM 后对磁链波动抑制效果较为明显，证明了该控制策略的可行性。

4 结语

本文将基于 SVPWM 的直接推力控制策略应用于初级永磁型直线电机控制系统，仿真结果表明，该控制系统运行平稳，动态响应速度快，系统抗干扰能力得到增强。相比于传统的直接推力控制系统，其对系统推力和磁链波动的抑制取得了较好效果。因此，对今后的进一步研究具有重要的指导意义。

参考文献：

[1] 杜怿，程明，邹国棠. 初级永磁型游标直线电机设计与静态特性分析[J]. 电工技术学报，2012，27(11)：22-30.

[2] 唐小利. 永磁同步直线电机的矢量控制系统研究[J]. 电子技术与软件工程，2017(1)：112-112.

[3] 肖鹏，王伟，张颖，等. 永磁同步直线电机矢量控制系统中初始寻相和电角度的测定[J]. 微电机，2009，42(11)：1-3.

[4] 陈修亮，车倍凯. 永磁同步电机矢量控制解耦方法的研究[J]. 电气技术，2013(4)：37-40.

[5] 杨俊友，崔皆凡，何国锋. 基于空间矢量调制和滑模变结构的永磁直线电机直接推力控制[J]. 电工技术学报，2007，22(6)：24-29.

[6] 王丽梅，程兴民. 改进的永磁直线同步电机直接推力控制[J]. 组合机床与自动化加工技术，2015(12)：60-64.

[7] 陈浩然，汪旭东，许孝卓，等. 改进的永磁同步直线电机直接推力控制策略[J]. 电子测量技术，2017，40(2)：38-41.

[8] 郎宝华，刘卫国，周熙炜. 基于参考磁链空间电压矢量调制策略的 PMSM DTC 系统[J]. 电气传动，2007，37(7)：21-25.

[9] 朱国昕. 基于 SVPWM 的永磁同步电动机直接转矩控制[J]. 变频器世界，2012(5)：64-66.

（责任编辑：黄 健）