文章编号：2096-1472(2021)-05-07-03
DOI:10.19644/j.cnki.issn2096-1472.2021.05.002

基于关联分析的中老年体检数据的挖掘

郭慧敏

(安徽大学经济学院，安徽 合肥 230601)
✉17755895356@163.com

摘　要：根据中老年体检报告，运用Apriori算法挖掘各个指标之间的联系，为医生、患者提供诊断参考与建议。通过安徽省某三甲医院的体检数据，筛选出40岁及以上的中老年人群为研究对象，应用数据挖掘中关联规则的Apriori算法对超重、心电图、脂肪肝、血脂、血压、血糖、尿常规、吸烟、饮酒、总胆固醇等体检指标之间的关联关系进行分析研究。研究表明，体检者的个人不良习惯、超重、高龄、高血糖和脂肪肝等都密切相关，互相影响，提出中老年人群应加强对慢性疾病的预防，保持良好的作息习惯等相关建议。

关键词：数据挖掘；关联分析；Apriori算法；中老年体检
中图分类号：TP181　**文献标识码**：A

Data Mining of Physical Examination for the Middle-aged and Elderly based on Association Analysis

GUO Huimin

(*School of Economics*, *Anhui University*, *Hefei* 230601, *China*)
✉17755895356@163.com

Abstract: This paper proposes to use Apriori algorithm to mine the links between various indicators in the medical examination report of middle-aged and elderly people, which provides diagnosis references and suggestions for doctors and patients. The middle-aged and elderly people aged 40 and above are selected as the research objects from the physical examination data of a Class A tertiary hospital in Anhui Province. Then, Apriori algorithm of association rules in data mining is used to analyze and study the correlation between physical examination indicators, such as overweight, electrocardiogram, fatty liver, blood lipids, blood pressure, blood sugar, urine routine, smoking, drinking, and total cholesterol. Research results show that personal bad habits, overweight, advanced age, high blood sugar, and fatty liver of physical examinees are closely related and affect each other. This paper proposes that middle-aged and elderly people should strengthen the prevention of chronic diseases and maintain good work and rest habits.

Keywords: data mining; association analysis; Apriori algorithm; middle-aged and elderly physical examination

1 引言(Introduction)

近年来，大部分医院在移动医疗兴起的形势下，都建立了数字化医疗信息系统和患者的电子信息健康档案[1]，医院内部积累了大量医疗相关的数据，使得医疗信息数字化程度越来越高[2]。医疗数据不仅与每个人的生活和生命健康息息相关，而且对疾病的诊治与医学研究具有重要价值。然而目前大部分医院只是简单地进行患者医疗数据的采集与存储，缺乏对它们进行深层次的分析与利用，如何快速有效地在海量的医疗数据中发现潜在的有价值的信息是一项重大挑战[1]。

关联规则挖掘作为数据挖掘领域重要的研究分支，是当前在发展过程中比较重要、实用的技术[3]。在医学领域中，通过关联规则发现疾病患者中医症状之间的关联关系和其他症状之间存在的规律性，能够根据这些规律分析病因，预测疾病的发展[4]。本文以医院数据系统中的体检数据为研究对象，利用关联规则的Apriori算法，将每个病人的症状及其他病情诊断信息看作是一种购物篮，然后对其进行挖掘分析[2]，为个人健康提供预警，为医疗诊断提供科学依据参考。

2 关联规则算法(The Apriori algorithm)

2.1 Apriori算法概述

关联分析是由R.Agrawal等人提出的一种简单实用的非

监督学习算法[5]，反映了事物之间的依赖或关联，试图找到数据集中隐含的或感兴趣的关系，其结果通常以频繁项集或关联规则的形式表示。最经典的案例就是“啤酒与尿布”。沃尔玛超市根据详细的原始交易信息来对顾客的购物行为进行数据挖掘，来了解顾客在其门店的购买习惯，适当地调整货架，增加购买行为。然而，挖掘出来的规则在实际中并不是都有指导意义，比如说，如果一个客户买了杯子，就会有40%的可能性买茶叶，但是我们不能依据这个就把杯子和茶叶放在一起出售，我们借助置信度和支持度这两个评估指标来对关联规则进行有价值的评估，设置最小的支持度和置信度使我们得到的关联规则具有一定的参考价值。

2.2 相关概念

Apriori算法是关联规则算法，是非常经典的一种数据挖掘的算法，应用十分广泛，可以较好地发现数据之间的隐藏规则。

(1)项和项集。项为交易数据集中的每一种商品，项集为项的集合。

(2)事务。事务为交易数据集中对应的每一条记录。

(3)关联规则。关联规则指的是在X出现的同时，Y也会出现，其中X、Y均是I的真子集，并且二者交集不为空。

(4)支持度。支持度计算公式为：

$$Support(X,Y)=p(X,Y)=\frac{Num(X,Y)}{Num(All)}$$

表示XY同时出现的概率占总数的概率，表示X和Y两个事件同时发生的概率。

(5)置信度。置信度计算公式为：

$$Confidence(X\rightarrow Y)=p(X\mid Y)=\frac{P(Y)}{P(XY)}$$

表示在Y出现的条件下X出现的条件概率。

(6)频繁项集。频繁项集是指支持度不低于最小支持度的阈值的项集。

(7)强规则。强规则是指同时满足最小支持度阈值和最小置信度阈值的规则。

2.3 Apriori算法基本步骤与实现

Apriori算法的过程主要分为两步[6]：根据支持度阈值找出所有的频繁项集；通过置信度阈值找出频繁项集中的强关联规则。Apriori算法的基本步骤如下：

(1)首先扫描所有的数据集D，产生候选1−项集的集合$C1$。

(2)由候选1−项集的集合$C1$根据最小支持度产生频繁1−项集的集合$L1$。

(3)对$k>1$，重复执行步骤(4)、(5)、(6)。

(4)由Lk执行连接和剪枝操作，产生候选$(k+1)$−项集的集合$C(k+1)$。

(5)根据最小支持度，由候选$(k+1)$−项集的集合$C(k+1)$，产生频繁$(k+1)$−项集的集合$L(k+1)$。

(6)若$L\neq\Phi$，则$k=k+1$，跳往步骤(4)，否则往下执行。

(7)根据最小置信度，由频繁项集产生强关联规则，程序结束。

设置好最小支持度阈值和最小置信度阈值之后，Apriori算法开始执行，扫描数据集首先产生频繁1项集，将得到的频繁1项集进行连接操作，再次扫描数据集D得到满足最小支持度阈值的频繁2项集，以此类推直到频繁k项集[7]。

算法流程图如图1所示。

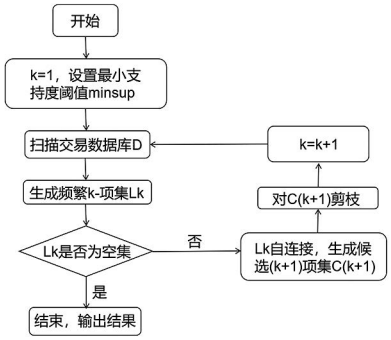

图1 算法流程图

Fig.1 Algorithm flow chart

3 数据来源与处理(Data source and processing)

3.1 数据来源与特点

本文的数据来源是安徽省某三甲医院2019年体检中心的体检数据，一共筛选2,345 份体检数据报告，采集的指标主要包含基本人物信息(性别、年龄等)以及相关的检查指标属性特征。

医疗数据的数据类型繁多。医疗数据包括纯数据、信号、CT、B超等医疗影像数据,文本类型有患者记录的身份、症状描述、检测和文本表示的诊断等多种模式，其数字类型有些是连续型数据，有些是离散型数据[8]，存在缺失值、异常值和重复值。收集到的医疗数据往往是不完整的,病人由于隐私不愿意透露或者由于人工记录的偏差、数据的不清晰表达、记录本身的不确定性等都造成了医疗信息的不完整性，且医院每天收集的大量数据可能会包含重复、无关紧要的数据[9]。数据中涉及个人的信息，如姓名、住址、身份证信息等，需要对隐私性、敏感性信息进行过滤。

3.2 数据预处理

数据预处理是进行数据挖掘必不可少的关键一步，目的是让数据适应模型，匹配模型的需求。数据预处理分为四个部分：数据清洗、数据集成、数据变换和数据归约[10]。

医疗原始信息包含体检人员的基本信息表和体检信息表，其中基本信息表包含姓名、住址、身份证号等一些敏感信息，这些涉及个人隐私的信息，需要进行脱敏处理；体检信息表包括血压、血脂、血糖等疾病情况，这些数据需要整理进行挖掘。数据清洗包括缺失值和异常值的处理。对于缺失值的处理，咨询相关医护人员或者查询相关病例记录进行空缺值填充，对于查询不到的缺失值用均值填补，异常值直接删除。数据集成是将多个数据源放在统一的仓库中，本文重点研究的是中老年人体检状况，筛选出40岁及以上的群体，针对其性别、年龄、高血压、高血脂、高血糖等检验指标信息之间的相关联性，剔除那些与研究不相关的属性记录，通过数据集成将相关表中需要研究的属性信息集成到一

个表中，将数据类型和数据单位进行统一化处理。数据变换是对数据进行规范化处理，本文中主要是数据离散化，进行关联分析。首先属性项不能是数值型的，像年龄、胆固醇水平等都是连续数值型数据类型，不能进行数据挖掘，将数据格式转换成英文或者数字化可以提高算法的运算效率，所以本文通过一定的标准把现有的文字数据格式进行英文字母、数字化或布尔值转换处理。这样做也是为了用关联规则更好地挖掘中老年群体病症之间的关系，满足数据挖掘的要求，比如年龄可以划分为两个年龄段：[40,65)、[65,max)，那么每个人的年龄就分别对应于相应的年龄段了，数值型数据变成离散化，其他几列连续数值型也是采用类似的方法离散化[11]。

因此，本文结合Apriori算法和医疗数据特点，查阅相关医学资料，对数据进行适当的离散化处理，将数据格式转换成事务性库，具体如表1所示。

表1 事务数据库映射表

Tab.1 Transaction database mapping table

字段名称	映射规则
性别	A1：男性　A2：女性
年龄	B1：[40,65) B2：[65,max)
体质指数	C：超重
心电图检查	D：心电图异常
脂肪肝检查	E：脂肪肝
血脂检查	F：高血脂
血压检查	G：高血压
血糖检查	H：高血糖
尿常规检查	I：尿常规异常
吸烟	J1：吸烟 J2：已戒烟
饮酒	K1：偶尔 K2：经常 K3：每天
总胆固醇	L：＞5.7 mmol/L
甘油三酯	M：＞1.7 mmol/L
高密度脂蛋白	N：＜1.0 mmol/L

得到事务项映射表之后，我们就可以利用该表得到具体需要挖掘的事务数据库D。扫描关系数据库中的数据表，对于每次扫描到的属性值，根据已经定好的事务项参照表，将该属性值所对应的具体编号写入事务表中，如表2所示。

表2 转换的事务逻辑表

Tab.2 Transformed transaction logic table

事务编号(TID)	项目集(ItemSet)
T1	{A2,B2,L}
T2	{A2,B2,C,D,E,F,L,H}
T3	{A2,B1,E,I,L}
T4	{A2,B1,C,E,F,G,L,M}
T5	{A2,B1,C,L,M}
T6	{A2,B2,C,E,L,M}

数据的预处理阶段已经完成，接下来用Apriori算法挖掘的事务数据库来进行关联规则的分析。

4　应用与实现(Application and implementation)

运用Python软件进行关联规则挖掘，设置的最小支持度为0.03，置信度为0.80，由此挖掘得到以下有意义的规则和相关参数，如表3所示。

表3 体检数据关联规则

Tab.3 Association rules of physical examination data

序号	置信度	支持度	规则
1	0.988505747	0.036673774	1C,J1⇒A
2	0.973684211	0.063113006	J1⇒A1
3	0.967741935	0.038379531	B2,J1⇒A1
4	0.941747573	0.041364606	K1⇒A1
5	0.900000000	0.030703625	K3⇒A1
6	0.867469880	0.030703625	A2,D,E,I⇒C
7	0.853211009	0.039658849	D,E,I⇒C
8	0.849315068	0.079317697	A2,D,E⇒C
9	0.845528455	0.044349680	B2,E,I⇒C
10	0.842696629	0.031982942	C,D,L⇒A2
11	0.841269841	0.045202559	B1,D,E⇒C
12	0.840579710	0.049466951	A2,B2,D,E⇒C
13	0.840455840	0.125799574	D,E⇒C
14	0.840000000	0.080597015	B2,D,E⇒C
15	0.839080460	0.031130064	A2,B2,E,I⇒C
16	0.839080460	0.031130064	A1,B2,D,E⇒C
17	0.834710744	0.043070362	B2,E,H⇒C
18	0.828358209	0.047334755	D,E,M⇒C
19	0.825757576	0.046481876	A1,D,E⇒C
20	0.821428571	0.068656716	A1,B2,E⇒C
21	0.821052632	0.033262260	A1,E,G⇒C
22	0.808383234	0.057569296	A2,E,I⇒C
23	0.806818182	0.030277186	E,I,M⇒C
24	0.804469274	0.061407249	B2,E,M⇒C
25	0.803493450	0.078464819	E,I⇒C
26	0.821052632	0.033262260	A1,E,G⇒C

本文给出了置信度为前26的排名。通过以上规则，在中老年人群中，我们可以得出以下结论：

(1)吸烟、体重超重，还经常喝酒的以中老年男性群体为主。

(2)针对老年人群，心电图异常、有脂肪肝，并且尿常规异常的，一般都体重超重。

(3)体重超重、心电图异常并且胆固醇较高的中老年女性居多。

(4)年龄在65岁以上的老年人中，血糖较高的人群体重一般超重。

(5)心电图异常、尿常规异常、有脂肪肝并且甘油三酯偏高的人群超重。

(6)中老年男性中，有脂肪肝和高血压的体重一般偏重。

针对老年人群，在大多数人的认知里，“三高”等一系列慢性病似乎已经成为这个年龄段的代表符号。从本论文的研究结果可以看出，在中老年人这一群体中，随着年龄的增加、生活方式的改变、基础代谢率的下降，由于缺乏运动、社交增多以及其他不良的饮食习惯等原因，使得肥胖的发生率增加，偏重的体质大概率会伴随高血脂和高血压等一系列不良后果，给中老年人的生活质量带来极大的影响。为了有效地避免这些病症，引导中老年人群建立健康的生活方式，通过合理的饮食、科学的营养搭配、适当的锻炼、良好的习

(下转第6页)