文章编号:1000-5471(2013)10-0001-06

# 基于关联规则分析治疗 2 型糖尿病临床用药规律①

曹锦梅[1], 凌 灿[1], 赵小龙[2], 凌 丽[3]

1. 新疆医科大学 高等职业技术学院,乌鲁木齐 830054;2. 新疆医科大学 医学工程技术学院,乌鲁木齐 830011;
3. 西南大学 信息中心,重庆 400715

**摘要**:通过对医院临床糖尿病电子处方原始资料进行处理后,得出病人的症状与使用相关药物的信息,进行数据挖掘分析,从不同年龄、性别、证型方面得出某症状与药物使用之间存在的规则以及隐藏在所使用药物与药物之间的深层规律,可以帮助医生在大量临床数据分析中归纳出辅助诊疗决策,为临床用药和基础研究提供思路.

**关 键 词**:数据挖掘;关联规则;糖尿病

**中图分类号**:TP392 **文献标志码**:A

## 1 前 沿

近几年来,2 型糖尿病在全球范围内呈逐年上升趋势,在一些大型医院,每天有上千人去看病、开药,每种病与医生所开药方存在某些内在的联系或规律. 糖尿病在长期临床实践和中医学理论的指导下,也积累了许多丰富的经验,而研究糖尿病某症状与西医药物使用之间存在的规则,以及隐藏在所使用药物与药物之间的深层规律的文献甚少. 深入研究患者体质和发病时的病机,正确运用药方,对糖尿病治疗和预防并发症的发生无疑具有积极的意义[1].

## 2 数据挖掘技术

### 2.1 SSAS

Microsoft SQL Server Analysis Services(SSAS)为商业智能应用程序提供了联机分析处理(OLAP)和数据挖掘功能. Analysis Services 可以设计、创建和管理多维结构(其中包含从关系数据库等其他数据源聚合的数据),并通过这种方式来支持 OLAP. 对于数据挖掘应用程序,Analysis Services 允许使用多种行业标准的数据挖掘算法来设计、创建和可视化基于其他数据源的数据挖掘模型,或者也可以创建自己的数据挖掘算法,将这些算法加入到 SQL Server 中,并提供给开发人员和客户使用.

### 2.2 数据挖掘

数据挖掘又称为数据库中的知识发现,简称 KDD(Knowledge Discovery in Database),它是从大量的、不完全的、模糊的、随机的数据中,提取隐含在其中的、人们事先不知道的、但又是潜在有价值的模式或规律等知识的复杂过程,是统计分析方法学的延伸和扩展. 它与传统数据分析的本质区别是数据挖掘是在没有明确假设的前提下去挖掘信息、发现知识,所得到的信息具有先前未知、有效和实用 3 个特征.

① 收稿日期:2013-05-02
基金项目:新疆医科大学科研创新基金课题(XJC2011113);中央高校基本科研项目(XDJK2013C142).
作者简介:曹锦梅(1976-),女,甘肃武威人,硕士,副教授,主要从事医学数据分析的研究工作.
通信作者:凌 丽.

### 2.3 关联规则分析

两个或多个变量的取值之间存在某种规律或数据对象之间的相互依赖关系称为关联. 关联分析的任务就是从数据中发现那些确信度和支持度都大于给定值的规则的关联. 算法采用的策略：首先找出所有频繁项集：其目标是发现满足最小支持度阈值的所有项集，这些项集为频繁项集. 然后从频繁项集产生强关联规则，目标是从频繁项集中提取所有高置信度的规则. 在进行挖掘时先设定数据的最小支持度(min_sup)和最小置信度(min_con)，只有满足最小支持度的项集(元素)才是频繁项集，在频繁项集基础上挖掘出来的规则满足最小置信度时，此关联规则才能成立.

在医学信息数据库中，建议引擎可以根据病人已经确诊的疾病和表征或病人在治疗过程中的反应，分析治疗方案实施过程中诱发其他疾病的概率以及与时间的关系等. Microsoft 关联算法可以在医学信息数据库模型中找到许多规则[3].

### 2.4 改进的 Apriori 算法

Apriori 算法是关联规则挖掘中的经典算法，由 Agrawal R 等人在 1993 年提出，该算法的核心是基于频繁集理论的递归方法，用于挖掘事务数据库中项集间的关联规则问题[4]. 这是一个基于两阶段频繁集的思想方法，将关联规则挖掘算法的设计分解为以下两个子问题：[2]

1) 找到所有支持度大于最小支持度的项集，这些集称为频繁集.

2) 使用第一个子问题找到的频繁集产生期望的规则.

它是一种逐层搜索的迭代方法，k 项集用于探索(k+1)项集，首先找到频繁 L -项集的集合，该集合记为 L1，L1 用于频繁 2 -项集的集合 L2，而 L2 用于找 L3，直到不能找到频繁 k -项集. 此算法在生成频繁项集时会产生大量的候选项集需要检验，并且每次进行频繁 k -项集判定时，都必须扫描整个数据库，时间、空间花费极大. 因此会出现挖掘效率低等问题，针对此问题进行算法的改进，改进后算法的性质：频繁项集的所有非空子集都必须也是频繁项集[2].

此性质说明：若有项集 B 出现的频率不满足阈值 min_sup，则 B 不是频繁的，即支持度 P(B)＜min_sup. 如果项集 A 并到 B 中，则结果项集(BUA)不可能比 B 更频繁出现，因此，BUA 也不可能是频繁的，即支持度 P(BUA)＜min_sup，利用这个性质可以来缩小搜寻空间. 在进行挖掘时要设定数据的 min_sup 和 min_con，只有满足 min_sup 的项集(元素)才是频繁项集，在频繁项集的基础上挖掘出来的规则应满足 min_con，此关联规则才能成立.

Apriori 算法存在多次扫描数据库可能产生大量候选及反复对候选项集和事务进行模式匹配的缺陷，导致算法效率较低. 为了提高算法效率，引入了修剪技术来减小候选集 Ck(k 表示候选集中的元素个数)大小，由此可以显著地改进所有频集算法的性能. 算法中引入的修剪策略基于这样一个性质：一个项集是频集当且仅当他的所有子集都是频集，那么如果 Ck 某个候选项集有一个(k－1)-子集不属于 L(k－1)(L 表示一个满足最小支持度的频繁项集)，则这个项集可以被修剪掉，不再被考虑. 这个修剪过程可以降低计算所有候选集支持度的代价，从而提高了算法的效率.

## 3 数据挖掘准备

获取某医院临床电子处方原始资料，包括门诊病人信息和用药信息，如表 1,2 所示. 对数据做预处理后得到糖尿病患者症状与患者使用相关药物的信息，通过 2 个数据源共有的属性 VISIT_NO 来建立起 2 个表的连接，用 SQL 的连接查询方法，找出糖尿病治疗药物，得到 csv 格式的数据源，然后用数据转换功能完成数据格式的转换，从而得到真正的数据源. 采用 SQL 2005 对数据进行关联分析，挖掘出糖尿病症状与药物使用之间存在的规则，以及隐藏在所使用药物与药物之间的深层规律，然后进行规则分析，症状关联规则本身就蕴含了很多有价值的信息，但有些规则本身并不能挖掘出任何有价值的信息，所以只有根据出现的不同症状才能更好辨别，研究出现症状关联规则的原因就可以得到症状间存在的深层规则.

**表 1　病人信息及疾病数据源**

| 编号 | 字段名称 | 字段说明 |
|---|---|---|
| 1 | VISIT_DATE | 问诊日期 |
| 2 | VISIT_NO | 病历号 |
| 3 | CLINIC_LAB | 挂号科室 |
| 4 | SEX | 性别 |
| 5 | AGE | 年龄 |
| 6 | SYMPTOM | 病症 |
| 7 | RESULT | 疾病 |

**表 2　药物信息数据源**

| 编号 | 字段名称 | 字段说明 |
|---|---|---|
| 1 | VISIT_NO | 病历号 |
| 2 | PHAM_NAME | 药物名 |
| 3 | PHAM_CODE | 药物编号 |
| 4 | PHAM_SPEC | 药物规格 |
| 5 | FIRM_ID | 生产厂家 |
| 6 | PHAM_AMOUNT | 服用药物量 |
| 7 | FREQUENCY | 服用频率 |

## 4　数据挖掘

SSAS 提供了用于数据挖掘的工具 BIDS(Bus*in*ess Intelligence Development Sudio)，采用这个工具完成数据挖掘.

### 4.1　创建数据源和数据源视图

新建数据源主要负责为源数据连接存储信息. 数据源视图是针对项目中选定数据源的表或视图的元数据，将元数据存储在数据源视图中，使用户可以在开发过程中脱离对数据源的连接而使用元数据[3]. 数据源视图包含着数据库相关子集信息，此信息不只局限于源数据库中表的物理结构，还可以添加诸如关系、表和列的名称、计算列和命名查询之类的信息[3]，创建一个基于上一步的主数据源，然后添加来自辅助数据源的表.

通过建立两张表属性 VISIT_NO 的连接，一张表嵌套到事例集中，能使某个糖尿病患者与其所用的治疗药物信息相一致，才能在后面的数据挖掘中挖掘出某个糖尿病患者其治疗药物之间存在的联系，从多个患者中就能通过数据培训找出的治疗药物间存在的规则，建立 VISIT_NO 连接.

### 4.2　创建数据挖掘模型结构

数据挖掘模型的结构主要由数据挖掘列和数据挖掘算法来定义，挖掘结构是从该项目现有数据源视图中派生的. 本文采用数据挖掘中“Microsoft 关联规则”算法，在这一步需要对 2 个表选定谁为事例和谁为嵌套，事例作为挖掘的目标，嵌套作为挖掘预测的对象. 关联规则算法使用事例集的列来定义，通过再提交给数据挖掘模型的培训数据上运行算法来生成预测模型. 但嵌套表与事例表必须是多对一的关系，否则将不能生成挖掘模型.

### 4.3　数据挖掘模型生成

生成的数据挖掘模型并不是预测数据模型，必须经过设置关联规则算法的参数，运行关联规则算法进行数据培训，得出来的规则、信息才组成预测挖掘模型. 在此设置最小支持度的阈值为 50，最小置信度的阈值为 0.3，其他按默认值设置. 在培训过程中数据必须保持一致性，即一个患者对应的治疗药物必须一致，否则将出现挖掘结果错误.

通过以上过程的实施，最终得到挖掘出来的数据. 数据挖掘的流程如图 1 所示.

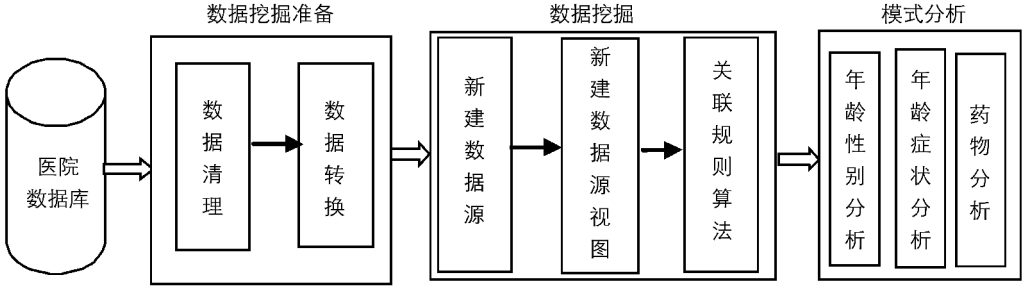


图 1　数据挖掘流程图

## 5 糖尿病临床用药结果分析

### 5.1 年龄、性别分析及结论

通过对1 435个糖尿病患者的性别、年龄进行关联规则分析发现以下强规则：AGE=(39，50)→SEX=男，发生的概率最高为0.65. AGE=(50，61)→SEX=女，发生的概率最高为0.56. 如表3所示：

从以上分析可得出，男人患糖尿病的年龄普遍要比女人患糖尿病的年龄低，并且在这2个年龄段中，数据挖掘结果的重要性也很高.

**表3 年龄、性别分析结果**

| 性别 | 关联规则 | 概率 |
|---|---|---|
| 男性 | AGE=(39，50) | 0.65 |
| 女性 | AGE=(50，61) | 0.56 |

### 5.2 年龄、症状分析及结论

年龄在45岁以下的主要症状如表4所示：

**表4 45岁以下症状分析**

| 置信度 | 支持度 | 作用度 | 统计数 | 结 论 |
|---|---|---|---|---|
| 78.13 | 30.86 | 1.22 | 50 | 口渴欲饮 & 夜尿频多→口干口渴 |
| 86.93 | 12.35 | 4.27 | 20 | 多梦→失眠 |

从表4中可看出，45岁以下的人主要症状为口干口渴，口渴欲饮和夜尿频多同时出现，与专业医生咨询可得出水道失调，开阀失职，则夜尿频多，热盛伤阴则口干口渴，口渴欲饮失眠伴有多梦，在此年龄主要考虑为阴虚阳盛，肝肾不足所致.

年龄在46～60岁的主要症状如表5所示：

**表5 46～60岁症状分析**

| 置信度 | 支持度 | 作用度 | 统计数 | 结 论 |
|---|---|---|---|---|
| 84.62 | 17.32 | 3.71 | 66 | 多梦 & 失眠→健忘 |
| 65.67 | 11.55 | 1.13 | 44 | 消瘦 & 体重下降→口干口渴 |
| 75.47 | 10.50 | 1.98 | 40 | 夜尿频多 & 消瘦 & 体重下降→夜尿频多 |

由表5可看出，此年龄段主要病机特点为：口干口渴、夜尿频多、体重下降、消瘦等消耗性症状组合出现，经咨询得出此类状况为热盛伤阴. 而对比45岁以下人群，夜尿频多的出现提示随着患者年龄的增大，肾虚表现日益突出. 失眠、多梦和健忘的成组出现是由热盛伤阴、肾阴亏虚、肝肾俱虚造成的.

年龄在60岁以上的主要症状如表6所示：

**表6 60岁以上症状分析**

| 置信度 | 支持度 | 作用度 | 统计数 | 结 论 |
|---|---|---|---|---|
| 97.96 | 21.33 | 1.21 | 96 | 夜尿频多 & 气短懒言 & 口干口渴→倦怠乏力 |
| 84.62 | 19.56 | 3.59 | 88 | 失眠 & 多梦→健忘 |
| 81.63 | 17.78 | 1.72 | 80 | 夜尿频多 & 气短懒言 & 口干口渴→夜尿频多 |
| 73.20 | 15.78 | 3.79 | 71 | 夜尿频多 & 失眠 & 健忘→眼前飞蝇 |
| 69.07 | 14.89 | 1.55 | 67 | 夜尿频多 & 失眠 & 健忘→视物模糊 |
| 61.64 | 10.00 | 1.59 | 45 | 眼前飞蝇 & 失眠 & 健忘→肢体麻木 |

由表6可看出，此年龄段主要病机特点为：夜尿频多、气短懒言、倦怠乏力、口干口渴伴随夜尿频多，同时出现失眠、健忘和多梦. 咨询专业医生可得出肝脾肾气阴两虚；失眠、健忘、夜尿频多伴随眼前飞蝇或视物模糊，提示肝肾气阴两虚；肾阴虚不能上济心火，心神不宁或肝不藏魂均可失眠，髓海不足则健忘；年老体衰，肾精亏虚，又肝肾同源，则肝血不足，目失所养，故眼前飞蝇；肝血不足或血运不畅，筋失所养，则肢体麻木.

结论：任何年龄、任何病程阶段的证侯表现特点均反映了肝脾肾证侯关联出现的规律，而表现在不同的年龄阶段、不同的病程阶段又各有侧重，从而得出治疗糖尿病必须“立足肝脾肾”三者同时治理才能有积极效果.

随着年龄的增加，糖尿病患者的证侯表现有实转虚的病变趋势，反应了年老体衰、脏腑机能减退、正

气自虚的自然规律. 随着病程的进展，证侯表现在病位上由肝脾肾三脏同病，而脾肝为主或脾肾为主逐渐向肝肾为主转变；在病性上由虚实夹杂而实证偏重向虚实夹杂而虚证偏重或因本虚渐甚而致标实渐重转化. 通过不同症状之间同时出现的概率的关联性，应用关联规则方法来筛选常见症状组合，是一种比较可行而有效的从大量数据中探索规律的方法，可以应用到证素研究和临床辨证规范研究中来.

### 5.3　药物分析结果

将数据挖掘的参数设置为 min_sup＝50，min_con＝0.3. 通过对糖尿病与其使用药物进行数据挖掘发现：糖尿病患者 1 435 人中，奥曲肽的支持度为 81，阿托伐他汀的支持度为 65. 糖尿病肾病治疗药物：奥曲肽药物与糖尿病患者的关联性最大，奥曲肽和糖尿病在满足最小支持度 50 和最小置信度 0.3 的情况下，事件糖尿病→奥曲肽发生的概率为 0.778，这是一个非常强的规则. 经过网上权威(中国糖尿病研究所)认证，奥曲肽是糖尿病肾病生长因子抑制剂，生长激素/胰岛素样生长因子(GH/IGFs)轴异常是发生糖尿病肾病的重要原因之一. 使用生长激素释放抑制剂奥曲肽(Octreotide)能抑制肾脏和肾小球肥大并降低蛋白尿，同时血糖控制不受影响，临床上已试用于糖尿病肾病的治疗，收到了良好的效果. 糖尿病→奥曲肽强规则如表 7 所示：

**表 7　药物分析—奥曲肽**

| 概率 | 重要性 | 规　　则 |
|---|---|---|
| 0.778 | 0.731 | 奥曲肽＝现有的－＞＃5％葡萄糖注射液(袋)＝现有的 |
| 0.778 | 0.731 | 奥曲肽＝现有的，RESULT＝糖尿病－＞＃5％葡萄糖注射液(袋)＝现有的 |
| 0.563 | 0.629 | ＃5％葡萄糖氯化钠注射液(袋)＝现有的，RESULT＝糖尿病－＞＃5％葡萄糖注射液(袋)… |

糖尿病并发高血脂治疗药物：阿托伐他汀与糖尿病的关联性也很高，在糖尿病和阿托伐他汀满足最小支持度 50、最小置信度 0.3 的前提下，事件(糖尿病，阿托伐他汀)→阿司匹林事件发生的概率为 0.308. 阿托伐他汀在治疗糖尿病中起着重要作用，它主要用于治疗糖尿病引起的并发高血脂，阿托伐他汀通过调脂作用降低 2 型糖尿病大血管并发症，其机制是通过竞争性的与合成胆固醇的限速酶结合，抑制此胆固醇合成的限速酶，增加肝细胞胆固醇受体表达，激活胆固醇受体活性，加速 LEL 的清除. 糖尿病→阿托伐他汀强规则如表 8 所示：

**表 8　药物分析—阿托伐他汀**

| 概率 | 重要性 | 规　　则 |
|---|---|---|
| 0.308 | 0.541 | 阿托伐他汀钙片(立普妥)＝现有的－＞＃阿司匹林肠溶片(拜阿司匹林)＝现有的 |
| 0.308 | 0.541 | 阿托伐他汀钙片(立普妥)＝现有的，RESULT＝糖尿病－＞＃阿司匹林肠溶片(拜阿司匹林) |

## 6　结　　论

针对糖尿病人的电子处方信息，从年龄、性别、证型 3 方面进行数据挖掘分析，得出男人患糖尿病的年龄普遍要比女人患糖尿病的年龄低；任何年龄、任何病程阶段的证侯表现特点均反映了肝脾肾证侯关联出现的规律，从而得出治疗糖尿病必须“立足肝脾肾”三者同时治理才能有积极的效果. 糖尿病肾病治疗药物：奥曲肽药物与糖尿病患者的关联性最大；糖尿病并发高血脂治疗药物：阿托伐他汀与糖尿病的关联性也很高，在治疗糖尿病中起着重要作用，主要用于治疗糖尿病引起的并发高血脂. 以上深层规律的发现，可以帮助医生从大量临床数据分析中归纳出辅助诊疗决策，为临床用药和基础研究提供一些思路.

**参考文献：**

[1]　赵丹丹. 数据挖掘在治疗糖尿病中药方剂数据库中的应用模拟 [D]. 青岛：中国海洋大学，2006.

[2]　陈文伟. 数据仓库与数据挖掘教程 [M]. 北京：清华大学出版社，2006.

[3]　李雯娟，曾照芳，陈　睿. 基于医学信息数据仓库模型的数据挖掘 [J]. 生物信息学，2009，7(2)：146－149.

[4]　姜黼莉，孟凡荣，月　勇. 多值属性关联规则挖掘的 Q-Apriori 算法 [J]. 计算机工程，2011，37(9)：81－83.

[5]　赵　霞. 基于数据挖掘技术的电子病历数据质量分析系统的研究与实现 [D]. 广州：华南理工大学，2010.

[6] MIAO Zhi-min, PAN Zhi-song, HU Gu-yu, et al. Treating Missing Data Processing Based on Neural Network and AdaBoost [C]. Proceedings of 2007 IEEE International Conference on Grey Systems and Intelligent Services, 2007: 1107－1111.

[7] ANDRITSOS P, MILLER R J, TSAPARAS P. Information Theoretic Tools for Mining Database Structure From Large Data Sets [C]. Proceedings of the 2004 ACM SIGMOD International Conference on Management of Data, 2004: 731－742.

[8] 周 怡, 王世伟, 彭 勇, 等. 医学数据挖掘—SQL Server 2005 案例分析 [M]. 北京: 中国铁道出版社, 2008: 24－126.

# On Analysis of Pharmaceutical Patterns in Clinical Treatment with Type II Diabetes Based on Association Rules

CAO Jing-mei[1], Ling Can[1], ZHAO Xiao-long[2], Ling Li[3]

1. *Vocational and Technical College, Xinjiang Medical University, Urumqi 830054, China;*

2. *Medical Engineering College, Xinjiang Medical University, Urumqi 830011, China;*

3. *Information Center of Southwest University, Chongqing 400715, China*

**Abstract**: By means of raw data processing of clinical diabetes electronic prescriptions, this paper draws the patient's symptoms and the use of related drugs, data mining analysis, in different ages, genders, and syndromes to get existing patterns and hidden deep principles between certain symptoms of drug use and the drug-taking patterns, which can help the doctor sum up in a large number of clinical data analysis the auxiliary treatment decisions, and to provide ideas for clinical and basic research.

**Key words**: data mining; association rules; diabetes

**责任编辑** 周仁惠