ISSN 1673-9418 CODEN JKYTA8
Journal of Frontiers of Computer Science and Technology
1673-9418/2009/03(02)-0188-10
DOI:10.3778/j.issn.1673-9418.2009.02.007

E-mail:fcst@public2.bta.net.cn
http://www.ceaj.org
Tel:+86-10-51616056

# 出生缺陷监测数据中的朴素干预规则挖掘*

张 悦[1+],唐常杰[1],李 川[1],朱 军[2],曾春秋[1],唐 良[1],刘显宾[1]
1. 四川大学 计算机学院,成都 610064
2. 中国出生缺陷监测中心 四川大学华西医学院,成都 610065

# Mining Naïve Intervention Rules in Birth Defect Data*

ZHANG Yue[1+], TANG Changjie[1], LI Chuan[1], ZHU Jun[2], ZENG Chunqiu[1], TANG Liang[1], LIU Xianbin[1]
1. College of Computer Science, Sichuan University, Chengdu 610064, China
2. Birth Defects Supervising Centre of Western China Medical School, Sichuan University, Chengdu 610065, China
+ Corresponding author: E-mail: zhangyue@cs.scu.edu.cn

**ZHANG Yue, TANG Changjie, LI Chuan, et al. Mining naïve intervention rules in birth defect data. Journal of Frontiers of Computer Science and Technology, 2009,3(2):188-197.**

**Abstract:** Mining naïve intervention rules from birth defects data is current hot topic concerned by both the birth defect research area and data mining fields. Taking birth defects data as background, this study aims the modeling of naïve intervention rules and discovering the possible causes of specific birth defects from the rough data. The main contributions include: Proposes a new concept called naïve intervention rule (NIR); Proposes and implements the algorithm to mine NIR; Conducts extensive experiments. The empirical result shows that the newly proposed algorithm successfully discovers the causes of prenatal birth defect, and provides evidences to suggest the redirection for intervention decision to reasonable degree.

**Key words:** naïve intervention rule; birth defect; delta

**摘 要:**出生缺陷干预规则挖掘是目前医学界和数据挖掘界共同关注的课题。以出生缺陷数据为背景,研究了朴素干预规则建模,并试图发现某些出生缺陷的可能致因。提出了朴素干预规则模型以及朴素干预规则挖掘算法。实验表明,提出的算法能有效挖掘出围产儿缺陷的致因,并为出生缺陷干预工程的政策制定提供致病因素的最佳状态调整方向。

* The National Natural Science Foundation of China under Grant No.60773169 (国家自然科学基金); the 11th Five Years Key Programs for Science & Technology Development of China under Grant No.2006BAI05A01 (国家"十一五"科技支撑计划).
Received 2008-08, Accepted 2008-10.

**关键词**:朴素干预规则;出生缺陷;变化量
**文献标识码**:A　**中图分类号**:TP311

## 1　引言

在市场调控、扩招与就业和疾病监控等领域,干预是保障稳定的必要手段。决策者需要了解在何种条件下,用多大的干预投入,可得到多大干预效益。干预规则研究旨在发现数据的动力学规律，回答以上问题,使得政府能节约干预投资,获得较大效益。有重大的社会意义和一定经济价值。

我国二十多年的出生缺陷监测数据表明,每年因新生儿出生缺陷引起的经济损失高达 200 多亿元。由于出生缺陷存在病种、地区、人群等差异,所以急需准确描述全国出生缺陷发生状况和动力学规律。

关联规则能够发现潜在的关联信息和知识,例如可以找出围产儿基本信息中与出生缺陷有强关联的属性,以事务间的关联的方式为决策提供帮助。但是关联规则前件属性各个状态的差异对规则后件的影响程度却被忽视。实践表明,简单的关联规则支持度和置信度在出生缺陷的解释上可能表现出一定的误导性，它并不估量前件与后件之间关联的实际强度,而只是给定了条件概率的估计[1]。一个新的思路是:把"状态转移"模式引入关联规则,在规则中描述前件属性与后件属性的变化量关系,体现出前件的变化引起后件变化的方向和幅度,使关联规则以反映最佳干预方向的方式提供决策建议。在这一思想指导下,本文工作如下:(1)建立了朴素干预规则模型;(2)对朴素干预规则提出了相应的挖掘算法;(3)通过实验证明了该方法的有效性和实用性。

## 2　朴素干预规则建模

### 2.1　建模思想

关联规则(association rule,AR)的概念和挖掘方法由 Agrawal 等于 1993 年在文献[2]中提出并获得了很大的成功。传统的关联规则 $X_1 \Rightarrow Y_1$ 仅仅反映了事务 $X_1$ 和 $Y_1$ 间的相互作用。制定政策的专家常常问:"如果把某项经济技术指标 $X_1$(例如投入经费)干预为 $X_2$,$Y_1$ 会有什么变化? 是变好还是变坏?"。这正是干预规则挖掘旨在解决的问题。

为清晰表达问题,在本文中施加干预(intervention)的入口属性常记为 $I_1$,$I_2$,…,评价(evaluation)干预效果的属性常记为 $E_1$,$E_2$,…等等。

**例 1**　考虑表 1 的关系型数据集 $D$。

Table 1　Example of relational data set

**表 1　关系型数据集举例**

| ID | $I$ | $E$ |
|---|---|---|
| 1 | $a$ | $m$ |
| 2 | $a$ | $m$ |
| 3 | $b$ | $m$ |
| 4 | $a$ | $n$ |
| 5 | $b$ | $n$ |
| 6 | $c$ | $n$ |
| 7 | $c$ | $n$ |

挖掘与事务 $E(m)$有关的规则,得到关联规则 AR1:

$$I(a) \Rightarrow E(m) \quad Sup=2/7, Conf=2/3 \qquad (AR1)$$

AR1 表明 $X$ 属性的 $a$ 状态与 $E$ 属性的 $m$ 状态有强关联。如果用户希望今后 $E$ 属性的 $m$ 值在数据库中所占比率降低,则会考虑对 $X$ 属性施加干预。但是 $X$ 属性有三个属性值 $a$、$b$ 和 $c$,为取得好效果,用户问:把 $I$ 属性干预为 $b$ 还是 $c$?

借用理论物理学家在相对论中的 "思想实验"方法,考察下列两个思想实验的后果:

(1)假定把 $I$ 属性值从 $a$ 干预为 $b$,有何宏观后果?

(2)假定把 $I$ 属性值从 $a$ 干预为 $c$,有何宏观后果?

为回答上述问题，继续挖掘 $X$ 属性的各个值与 $Y$ 属性的 $m$ 值的关联关系,得到两条关联规则:

$I(b)\Rightarrow E(m)Sup=1/7,Conf=1/2$ (AR2)

$I(c)\Rightarrow E(m)Sup=0,Conf=0$ (AR3)

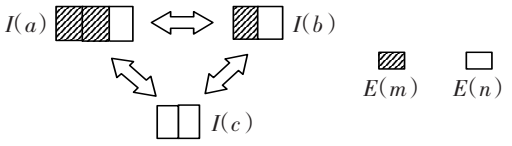


Fig.1 $I$'s each value and relative value of $E$

**图 1 数据 $D$ 的 $I$ 属性和 $E$ 属性各值分布**

规则 AR3 表明,当属性 $X$ 取值为 $c$ 时,$Y$ 的取值为 $m$ 的元组所占百分比最小。

综合 AR1~AR3,得到干预策略:把 $I$ 的值由 $a$ 干预为 $c$,可能达到降低 $m$ 出现率的最佳效果。

为形式化描述例 1 的核心思想,引入朴素干预规则(naïve intervention rule,NIR),表述为:

$I(a\rightarrow c)\Rightarrow E(m\rightarrow n)Sup=2/7,Delta=2/3,Conf=1$ (NIR1)

规则 NIR1 表示:$I$ 的值为 $a$ 且 $E$ 的值为 $m$ 的数据量占总数的 2/7,$E$ 为 $m$ 的数据量所占比例,在 $I$ 为 $a$ 的元组中比在 $X$ 为 $c$ 的元组中高出 2/3(=2/3-0),高出的部分占 $I$ 为 $a$ 的元组中 $E$ 值为 $m$ 的数据量的 100%。

朴素干预规则和传统关联规则的最大区别是前者引入了干预(或强制变化)概念。NIR1 中的 $I(a\rightarrow c)$ 的意义是"把 $I$ 的值由 $a$ 干预为 $c$"

NIR1 为决策者提供了明确的建议:如把 $I$ 的值由 $a$ 干预为 $c$,则总数据中,可能会有 2/7 的数据受此政策影响,$I$ 为 $a$ 的数据中,可能 $E$ 值为 $n$ 的数据量会增加 2/3,而这些数据全部由原来 $I$ 为 $a$ 且 $E$ 为 $m$ 的数据转变而来。

把这个例子形式化,下面将引入朴素干预规则模型。

## 2.2 朴素干预规则模型

挖掘干预规则时,需要从关系数据库派生出事务数据库。文献[3-4]中提供了布尔派生方法。对每个属性,按其各个属性值转换为项,以布尔值表示该事务是否发生。例如,把表 1 的关系型数据进行布尔转换,如表 2 所示。其中,1 表示发生,0 表示不发生。按表 2 生成事务数据库,如表 3 所示。

Table 2 Boolean transformation of table 1

**表 2 表 1 的布尔转换**

| ID | $I_1$:$I(a)$ | $I_2$:$I(b)$ | $I_3$:$I(c)$ | $I_4$:$E(m)$ | $I_5$:$E(n)$ |
|---|---|---|---|---|---|
| 1 | 1 | 0 | 0 | 1 | 0 |
| 2 | 1 | 0 | 0 | 1 | 0 |
| 3 | 0 | 1 | 0 | 1 | 0 |
| 4 | 1 | 0 | 0 | 0 | 1 |
| 5 | 0 | 1 | 0 | 0 | 1 |
| 6 | 0 | 0 | 1 | 0 | 1 |
| 7 | 0 | 0 | 1 | 0 | 1 |

Table 3 Transaction database of table 2

**表 3 表 2 的事务数据库形式**

| 事务 ID | 项集列表 |
|---|---|
| 1 | $I_1,I_4$ |
| 2 | $I_1,I_4$ |
| 3 | $I_2,I_4$ |
| 4 | $I_1,I_5$ |
| 5 | $I_2,I_5$ |
| 6 | $I_3,I_5$ |
| 7 | $I_3,I_5$ |

注意,表 3 中的 $I_1$、$I_2$ 和 $I_3$ 分别表示属性 $X$ 取值 $a$、$b$、$c$ 的事件,在一事务中,$I_1$、$I_2$ 和 $I_3$ 至多一个为真。这样派生出来的项称为互斥项。

在同一个互斥项集中,各个项互为彼此的互斥项。例如对"性别"属性转换得到的项"性别为男"和项"性别为女"是互斥项。互斥项集的大小为该属性的属性值个数,事务数据库中每个事务中的项集大小为关系数据库中的属性个数。基于互斥项,引入朴素干预规则概念:

**定义 1**(朴素干预规则)设数据集 $D$ 中用户选定的干预属性集 $I=\{I_1,I_2,\cdots,I_w\}$,用户选定的评价干预效果的属性集 $E=\{E_1,E_2,\cdots,E_n\}$,$I\cap E=\varnothing$。在 $I$ 上派生的所有项的集合记为 $I$,在 $E$ 上转换得到的所有项的集合记为 $E$,满足下列条件的表达式称为 $r$ 朴素干预规则:

(1)存在频繁项集 $L'\subseteq I$,对所有 $I_i'\in L'$,存在

$I_q' \in E$,使得关联规则 $r'$:$I_1' \wedge I_2' \wedge \cdots \wedge I_{|L'|}' \Rightarrow I_q'$ 为强关联规则,支持度为 $Sup(r')$,置信度为 $Conf(r')$。

(2)对每个 $I_i' \in L'$,设 $I_i'$的属性为 $I_i$,值为 $a_i$,存在 $I_i'$的互斥项 $I_i''$,属性值为 $a_j$,存在关联规则 $r''$:$I_1'' \wedge I_2'' \wedge \cdots \wedge I_{|L'|}'' \Rightarrow I_q$,支持度为 $Sup(r'')$,置信度为 $Conf(r'')$。

(3)项 $I_i(a_i \rightarrow a_j)$表示为 $I_i'''$,设 $I_q'$的属性为 $E_k$,属性值为 $d$,则项 $E_k(d \rightarrow \sim d)$表示为 $I_q'''$,则 $r$:$I_1''' \wedge I_2''' \wedge \cdots \wedge I_{|L'|}''' \Rightarrow I_q'''$。

(4)规则 $r$ 的支持度 $Sup(r)$、变化度 $Delta(r)$和置信度 $Conf(r)$的定义为:

$$Sup=Sup(r') \tag{1}$$

$$Delta(r)=Conf(r')-Conf(r'') \text{ 且 } Delta(r)>0$$

$$Conf(r)=\frac{Delta(r)}{Conf(r')}$$

按照朴素干预规则进行干预,支持度是干预所影响的范围大小的体现,是干预必要性的度量。变化量 $delta$ 表明 $a_i$状态被干预为 $a_j$状态后,原先为 $a_i$状态的元组中,有多少比例的元组的 $d$ 值消失了,是干预结果的效果大小的体现;变化量和支持度共同体现了该规则的重要性。置信度表明了变化量占原始状态情况的百分比,是变化准确率的体现。

在实践中,观察到出生缺陷数据有下列和普通经济类数据不同的特点:

**观察 1** 出生缺陷研究主要针对缺陷的诱因,所以频繁项集中必包含"有缺陷"项,且频繁项集转换规则时,只转换出一条后件为"有缺陷"的规则。在普通数据中,如果频繁项集大小为 $n$,则需要考察的规则数目有 $C_n^1+C_n^2+\cdots+C_n^{n-1}$。

**观察 2** 干预规则后件的"有缺陷"互斥项为"无缺陷",因此仅需要考察无缺陷的对比,干预规则后件只为:缺陷(有→无)。

**观察 3** 出生缺陷自变属性中,许多属性只有两种取值状态,例如"近亲结婚"的状态为"是"或"否","家庭先天疾病"的状态为"有"或"无"。当一个属性只有两个互斥项时,只需要计算一种状态转移即可,而在普通数据中,如果一个属性有 $n$ 个互斥项,则一个项需要计算 $n-1$ 种状态转移。

**观察 4** 如果规则前件各属性只有两个互斥项,根据文献[5]可知:$r'$和 $r''$互为彼此的负关联规则。关联规则 $X \Rightarrow Y$ 的负关联规则 $\sim X \Rightarrow Y$ 的其支持度 $Sup(\sim XY)=Sup(Y)-Sup(X \cup Y)$,置信度 $Conf(\sim XY)=(Sup(X)-Sup(X \cup Y))/(1-Sup(X))$[6]。由此可以计算得:当 $|L'|=2$ 时,定义 1 里的 $Delta(r)=(Sup(r)-Sup(X)^2)/(Sup(X)-Sup(X)^2)$,其中 $X=I_1' \wedge I_2' \wedge \cdots \wedge I_{|L'|}'$。这样,不用统计因变属性的末状态关联规则也可以进行计算了。

上述观察提示:干预规则的频繁项集挖掘更具目的性,后件可直接确定为"缺陷(有→无)",由频繁项集转换的规则少,出生缺陷属性状态转换简单,且挖掘中许多互斥项的计算不用再统计因变属性的末状态关联规则。这些情况,减少了挖掘的工作量。

## 2.3　朴素干预规则相关度量分析

关联规则使用相关度量来判定关联规则是否有趣。如果项集 $A$ 和 $B$ 正相关,则说明当事件 $A$ 出现得越多时,事件 $B$ 也出现得越多;如果项集 $A$ 和 $B$ 不相关,则说明事件 $B$ 的出现与事件 $A$ 无关;如果项集 $A$ 和 $B$ 负相关,则说明当事件 $A$ 出现得越多时,事件 $B$ 出现得越少。根据这一性质,得到朴素干预规则 $A_1 \rightarrow A_2 \Rightarrow B \rightarrow \sim B$ 的前件与后件的相关度量公式:

$$corr_{(A_1 \rightarrow A_2),(B \rightarrow \sim B)}=Delta(A_1 \rightarrow A_2 \Rightarrow B \rightarrow \sim B) \tag{2}$$

即:

$$corr_{(A_1 \rightarrow A_2),(B \rightarrow \sim B)}=\frac{P(A_1 \cup B)}{P(A_1)}-\frac{P(A_2 \cup B)}{P(A_2)} \tag{3}$$

公式(2)和(3)中,$A_1$ 表示初始项集集合,$A_2$ 表示互斥项集合,$B$ 表示干预效果项集合。

**定理 1** 设 $A_1 \rightarrow A_2 \Rightarrow B \rightarrow \sim B$ 为朴素干预规则,前件与后件的相关度量为 $corr_{(A_1 \rightarrow A_2),(B \rightarrow \sim B)}=\frac{P(A_1 \cup B)}{P(A_1)}-\frac{P(A_2 \cup B)}{P(A_2)}$,则:

(1)当 $corr_{(A_1 \rightarrow A_2),(B \rightarrow \sim B)}>0$,前件和后件正相关;

(2)当 $corr_{(A_1 \rightarrow A_2),(B \rightarrow \sim B)}=0$,前件和后件不相关;

(3)当 $corr_{(A_1\to A_2),(B\to\sim B)}<0$,前件和后件负相关。

**证明** 由参考文献[7]可知,判断规则 $A\Rightarrow B$ 的前件 $A$ 的出现对后件出现的提升度 $lift(A,B)=\frac{P(A\cup B)}{P(A)P(B)}$,则:

$$corr_{(A_1\to A_2),(B\to\sim B)}=P(B)(lift(A_1,B)-lift(A_2,B)) \quad (4)$$

(1)当 $corr_{(A_1\to A_2),(B\to\sim B)}>0$ 时,$lift(A_1,B)>lift(A_2,B)$,$A_2$的出现比 $A_1$的出现对“提升”$B$ 的程度更小,如果把 $A_1$转换为 $A_2$,则降低了 $B$ 的出现频率,因此干预规则前件和后件正相关;

(2)当 $corr_{(A_1\to A_2),(B\to\sim B)}=0$ 时,$lift(A_1,B)=lift(A_2,B)$,$A_1$的出现和 $A_2$的出现对“提升”$B$ 的程度相同,如果把 $A_1$转换为 $A_2$,$B$ 的出现频率不变,因此干预规则前件和后件不相关;

(3)当 $corr_{(A_1\to A_2),(B\to\sim B)}<0$ 时,$lift(A_1,B)<lift(A_2,B)$,$A_2$的出现比 $A_1$的出现对“提升”$B$ 的程度更大,如果把 $A_1$转换为 $A_2$,则提高了 $B$ 的出现频率,因此干预规则前件和后件是负相关。 □

从公式(4)可以看出,朴素干预规则的相关度量为 $A_1$和 $A_2$的出现对 $B$ 出现率的提升程度之差乘以 $B$ 的出现率,这正是朴素干预规则的“变化量”意义的体现。由公式(2)和(4)可得:

$$Delta(A_1\to A_2\Rightarrow B\to\sim B)=P(B)(lift(A_1,B)-lift(A_2,B)) \quad (5)$$

**推论 1** 设 $A_1\to A_2\Rightarrow B\to\sim B$ 为朴素干预规则,前件与后件的相关度量为 $corr_{(A_1\to A_2),(B\to\sim B)}=\frac{P(A_1\cup B)}{P(A_1)}-\frac{P(A_2\cup B)}{P(A_2)}$ 则:

$$corr_{(A_1\to A_2),(B\to\sim B)}=-corr_{(A_2\to A_1),(B\to\sim B)}$$

**证明**

$$corr_{(A_1\to A_2),(B\to\sim B)}=-[\frac{P(A_2\cup B)}{P(A_2)}-\frac{P(A_1\cup B)}{P(A_1)}]=-corr_{(A_2\to A_1),(B\to\sim B)}$$

定理 1 和推论 1 都说明,$A_1\to A_2\Rightarrow B\to\sim B$ 和 $A_2\to A_1\Rightarrow B\to\sim B$ 不会同时作为有趣规则。在实际挖掘中,当发现其中一条规则为负相关规则时,交换干预规则前件的项的干预顺序,就可以得到对应的另一条有趣的正相关干预规则。 □

**推论 2** 干预规则 $A_1\to A_2\Rightarrow B\to\sim B$ 的前件和后件在以下情况下恒为正相关:(1)$A_1$和 $B$ 正相关且 $A_2$和 $B$ 负相关;(2)$A_1$和 $B$ 正相关且 $A_2$和 $B$ 不相关;(3)$A_1$和 $B$ 不相关且 $A_2$和 $B$ 负相关。

**证明** (1)因为 $A_1$和 $B$ 正相关,所以 $lift(A_1,B)>1$,因为 $A_2$和 $B$ 负相关,所以 $lift(A_2,B)<1$,由公式(4)得 $corr_{(A_1\to A_2),(B\to\sim B)}>0$,由定理 1 可知干预规则 $A_1\to A_2\Rightarrow B\to\sim B$ 的前件和后件为正相关。同理可证(2)和(3)。 □

推论 2 表明,控制正关联规则的前件数量,而增加负关联规则的前件,可以达到有效的干预效果。

**推论 3** 干预规则 $A_1\to A_2\Rightarrow B\to\sim B$ 在以下情况下恒为负相关:(1)$A_1$和 $B$ 负相关且 $A_2$和 $B$ 正相关;(2)$A_1$和 $B$ 不相关且 $A_2$和 $B$ 正相关;(3)$A_1$和 $B$ 负相关且 $A_2$和 $B$ 不相关。

**证明** (1)因为 $A_1$和 $B$ 负相关,所以 $lift(A_1,B)<1$,因为 $A_2$和 $B$ 正相关,所以 $lift(A_2,B)>1$,由公式(4)得 $corr_{(A_1\to A_2),(B\to\sim B)}<0$,由定理 1 可知干预规则 $A_1\to A_2\Rightarrow B\to\sim B$ 的前件和后件为负相关。同理可证(2)和(3)。 □

## 3 朴素干预规则挖掘算法

朴素干预规则的挖掘算法由三个部分组成:

(1)把出生缺陷关系型数据库转换为事务数据库;

(2)挖掘频繁项集。规则后件的项会进行人工指定。

(3)对频繁项集里每个项计算互斥项的关联规则,挖掘出 Delta 值最大的朴素干预规则。

第(1)部分的实现方法,已经在本文第 2 章说明;对于第(2)部分,已有许多快速的算法可直接使用,如算法 Apriori[8]、FP-growth[9]等,并且如果设 $X\Rightarrow Y$ 是一个合法的规则,则添加右部一项限制条件以后,会有 $O(2^{|X|+|Y|-1})$个无用规则被删除[3];因此,第(3)部分是朴

素干预规则挖掘的核心问题,本文将重点讨论此部分。在算法 1 中假定已经求得了频繁项集并保存在集合 $L$ 中。

**算法 1** 挖掘朴素干预规则。

输入:频繁项集 $L$,最小支持度 $min\_sup$,最小变化度 $min\_delta$,最小置信度 $min\_conf$;

输出:朴素干预规则集 $NIR$。

过程:

(1)$NIR=\emptyset$;
(2)generate all rules $R1:X\Rightarrow Y$ in $L$;
(3)for each $R1$ by $L$
(4) if $Sup(R1)\geqslant min\_sup$ and $Conf(R1)\geqslant min\_conf$ then
(5)  for each frequent item $I'$ in$(L-Y)$
(6)  for each mutex item $I''$
(7)  generate rule $R2:X'\Rightarrow Y$;
(8)  $delta=Conf(R1)-Conf(R2)$;
(9)  $conf=delta/Conf(R1)$;
(10)  if $delta\geqslant min\_delta$ and $conf\geqslant min\_conf$ then
(11)  $NIR=NIR\cup\{X\rightarrow X'\Rightarrow Y\rightarrow\sim Y\}$;
(12)  end
(13) end
(14)return $NIR$;

算法 1 的关键描述是第(5)行到第(12)行,该步骤对规则 $R1$ 前件各频繁项集的互斥项集计算 $Conf(R2)$,以找到符合要求的朴素干预规则。该算法的时间复杂度为 $O(n^2)$。

算法 1 第(8)行中,需要 $Conf(R2)$,这可以通过再一次扫描数据而得到。但是当数据量很大时,数据保存在数据库或者文本文件中,反复扫描数据会导致系统效率低下。因此,在生成所有关联规则后,可以扫描一次数据,记录下每个属性的互斥项组合对 $Y$ 的置信度,以后需要 $Conf(R2)$时进行查阅就可以了,不用再扫描数据。算法 2 是对此方法的描述,可以把算法 2 插入到算法 1 的第(2)行和第(3)行之间。

**算法 2** 统计 AR 各互斥项组合与后件的置信度。

输入:关联规则集 $R1=\{X\Rightarrow Y|\ X\cup Y=L\}$

输出:置信度记录 $ConfR$。

过程:

(1)set record set $IDX=\emptyset$;
(2)for each rule in $R1$
(3) for each item $I'$ from 1 to $|L'|$ and it's mutex item $I''$
(4)  $IDX=IDX\cup\{I_1''\ I_2''\ \cdots\ I_{|L'|}''\ I_q'\}$;
(5) for each transaction $t$ in dataset
(6)  for each $idx$ in $IDX$
(7)  if $t$ & $idx==idx$ then
(8)  $ConfR[idx]$++;
(9)  end
(10)return $ConfR$;

算法 2 的第(3)行把各属性的互斥项进行组合并列出了需要统计的事务类型,在第(4)行,它们被保存在集合 IDX 中。在第(7)行,使用布尔按位与的方法检查该行数据中的项集是否包含了记录中要求统计的项集 $idx$,如果是,则该 $idx$ 的统计数据增 1。算法 2 使用了稍多的内存,但提高了效率。算法 1 可以通过查询置信度记录 $ConfR$ 获得 $Conf(R2)$。

算法 1 和算法 2 可以应用于普通数据。由第 2.2 节的观察可知,在出生缺陷检测数据上进行朴素干预挖掘比在普通数据上进行挖掘更加简单:规则后件已固定,且大多数因变属性的互斥项仅有两个。假设所有因变属性均只有两个值,算法 1 可以简化为算法3。

**算法 3** 挖掘出生缺陷朴素干预规则简化算法。

输入:频繁项集 $L$,最小支持度 $min\_sup$,最小变化度 $min\_delta$,最小置信度 $min\_conf$;

输出:朴素干预规则集 $NIR$。

过程:

(1)$NIR=\emptyset$;
(2)generate rule $R1:X\Rightarrow Y$ in $L$;
(3)if $Sup(R1)\geqslant min\_sup$ and $Conf(R1)\geqslant min\_conf$ then
(4) $delta=(Sup(R1)-Sup(X)^2)/(Sup(X)-Sup(X)^2)$;
(5) $conf=delta/Conf(R1)$;
(6) if $delta\geqslant min\_delta$ and $conf\geqslant min\_conf$ then
(7)  $NIR=NIR\cup\{X\rightarrow\sim X\Rightarrow Y\rightarrow\sim Y\}$;
(8) end
(9) end
(10)return $NIR$;

算法 3 第(2)行的 $Y$ 为"缺陷(有→无)",$X$ 为 $L$ 中除 $Y$ 外的项的并。这一步是从频繁项集产生关联规则。算法 3 比算法 1 少产生 $C_{|L|}^1+C_{|L|}^2+\cdots+C_{|L|}^{|L|-1}-1$ 条

规则。对于互斥项,根据观察 4 可以直接计算 *delta*,不用再扫描数据或者使用置信度记录。该算法的时间复杂度降为 O($n$)。

在出生缺陷检测数据中,数据库属性不超过 24 个,可被选做自变量属性的不超过 20 个。大多数属性的取值只有两个,例如"近亲结婚"的"是"和"否",但是个别属性值不止两个,具有最多取值的属性为"职业",共有 8 个属性值。对于值多于两个的属性,仍然采用算法 1 的第(5)到(12)行的方法进行计算。因此,算法 3 在出生缺陷数据上进行运行是可行的。

## 4 实验和性能分析

### 4.1 实验数据与预处理

数据来源:1986~1991 年的全国围产儿数据,其中 1986 年和 1987 年的数据为围产儿基本信息数据,1988~1991 年的数据为"正常-缺陷儿"对照表数据和月统计报表数据。对照表数据和月统计报表数据,按数据制作所需的围产儿基本信息数据库,作为挖掘对象。数据量为 200 万条,围产儿数据有 24 个属性。

数据预处理:所有数据均以关系数据格式保存,做了预处理,根据出生缺陷数据的特点,数据集要转换成包含项的事务数据格式。包括下列步骤:

(1)把数值类型离散化。对容易人为划分的的数据,采用医师已经分割好的区间,融入医生的专业知识,例如:划分体重为偏瘦、正常、超重[3];对于其他类型,采用基于熵的方法寻找分割点。

(2)使用文献[3-4]的方法,将关系数据库中的多值、量化属性进行布尔转换,转换为事务数据库。

出生缺陷模型关注缺陷的有无,因此,关联规则后件可以人为指定,例如只表达新生儿是否有缺陷。这也使得频繁项集的挖掘更加简单。

### 4.2 实验结果与性能分析

实验平台:AMD Athlon 1.61 GHz 和 512 M 内存的 Windows XP 平台,Borland C++ Builder 6.0。

**实验 1** 参数:最小支持度为 0.000 5,最小变化度绝对值为 0.005 0,最小置信度为 0.200 0,限定挖掘范围为 1986 年和 1987 年的数据,选择全部可选属性为自变属性,限制规则后件为"缺陷儿(是→否)",获得规则集合 1,其中部分规则如表 4 所示。规则 0 到 2 表示"既往病史"、"近亲结婚"、"先天患病"和出生缺陷率有一定规律,并推荐干预为"无"。规则 3 和 4 表示"文化程度"为"文盲"和"硕士"时,缺陷率较高。规则 3 推荐把文化程度干预成"小学到大学",可能会有 0.91%的人不会出现缺陷,规则 4 推荐干预成"其他"的文化程度,可能会有 4.04%的人不会出现缺陷。规则 5 表示"母亲年龄"在 20 岁前和 36 岁后产生的围产儿更多,推荐产妇年龄在 20 岁到 36 岁之间。

Table 4 Some of results in the first experiment

**表 4 实验 1 的部分结果**

| ID | NIRule |
|---|---|
| 0 | 时间(1986~1987)∧既往病史(有→无)⇒缺陷儿(是→否) *Sup*=0.002 0,*Delta*=0.040 0,*Conf*=0.542 2 |
| 1 | 时间(1986~1987)∧近亲结婚(父母→无)⇒缺陷儿(是→否)*Sup*=0.024 0,*Delta*=0.338 8,*Conf*=0.990 6 |
| 2 | 时间(1986~1987)∧先天患病(有→无)⇒缺陷儿(是→否) *Sup*=0.000 5,*Delta*=0.462 2,*Conf*=0.930 7 |
| 3 | 时间(1986~1987)∧文化程度([1,1]∪[6,6]→[2,5])⇒缺陷儿(是→否) *Sup*=0.002 3,*Delta*=0.009 1,*Conf*=0.208 7 |
| 4 | 时间(1986~1987)∧文化程度([1,1]∪[6,6]→[7,7])⇒缺陷儿(是→否) *Sup*=0.002 3,*Delta*=0.040 4,*Conf*=0.926 9 |
| 5 | 时间(1986~1987)∧母亲年龄([17,20]∪[36,45]→[20,36])⇒缺陷儿(是→否) *Sup*=0.000 1,*Delta*=0.006 1,*Conf*=0.149 2 |

**实验 2** 参数:最小支持度为 0.000 1,最小变化度绝对值为 0.005 0,最小置信度为 0.100 0,使用 1986~1991 年的全部数据,选择参与挖掘的属性有"年龄"、"城乡"、"性别"、"胎数"、"出生缺陷"和"围产儿死亡",限制规则后件为"缺陷儿(是→否)"和"围产儿死亡(是→否)",获得规则集合 2,集合 2 包含集合 1,设定较高的朴素干预规则选取阈值可以得到更少、更有普遍意义的规则。集合 2 除去集合 1 内容外的部分规则如表 5 所示。规则 6 表示:(1)农村的缺陷儿死亡率较城镇的高。(2)若要降低死亡率,推荐干预为城镇的非缺陷儿。此外,由于城乡地域在现实生活中是无法干预的,则该规则只能起到发现相关性的作用。规则

7 和规则 5 类似。

Table 5　Some of results in the second experiment

**表 5　实验 2 的结果**

| ID | NIRule |
|---|---|
| 6 | 城乡(乡村→城镇)∧缺陷儿(是→否)⇒围产儿死亡(是→否) *Sup*=0.011 8,*Delta*=0.828 4,*Conf*=0.983 2 |
| 7 | 母亲年龄([17,20]∪[36,45]→[20,36])⇒缺陷儿(是→否)*Sup*=0.000 1,*Delta*=0.005 9,*Conf*=0.145 7 |

表 5 中两个规则集合表明,新方法可以挖掘出产妇各关键属性的状态转移情况对新生儿缺陷情况的影响程度。这对分析诱发新生儿缺陷的因素提供了难能可贵的素材,为决策者制定出生缺陷干预措施方案提供了理论基础。在实验中,仍然存在一些无用的规则,有些是由于属性连续值划分不合理引起的,有些是因为属性的不可干预特性引起的。对于前者,文献[7]中已经给出一个项分组的思想:许多项要经过医疗专家分组以消除无趣的关联,这个方面还有待于以后做更深入的研究。对于后者,可以人工选择有提示意义的进行保留,但是无法推荐干预方向。

图 2 展示了在出生缺陷统计数据上设定不同的最小支持度所得到的规则数量。可见,当最小支持度增大时,所能挖掘到的规则数量减少,当最小支持度增达到 0.02 时,规则数量非常少,并且逐渐趋近于 0。图 3 显示了不同的最小支持度下,在不同的数据量上进行挖掘所得到的规则数量。最小支持度越高,规则数量越少。随数据量的增大,规则数量趋于稳定。图 4 显示了在不同的最小支持度下,在出生缺陷数据上选

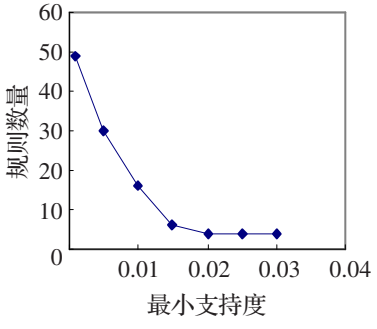


Fig.2　Number of rules by different *min_sup*

**图 2　最小支持度与规则数量变化图**

取不同的属性量时,算法第三部分的运行时间。图 5 显示了在不同的最小支持度下,在不同的数据量上算法第三部分的运行时间。从这两个图可知,最小支持度越高,运行时间越短。运行时间的增加速度随属性数量的增加而逐渐变大,而数据量和运行时间基本为线性关系,算法第三部分的伸缩性较好。

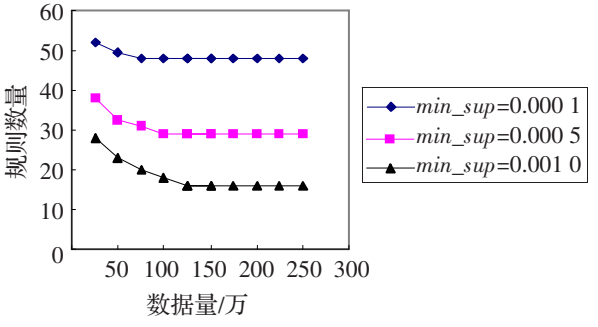


Fig.3　Number of rules by different data size

**图 3　数据量和规则数量的变化关系**

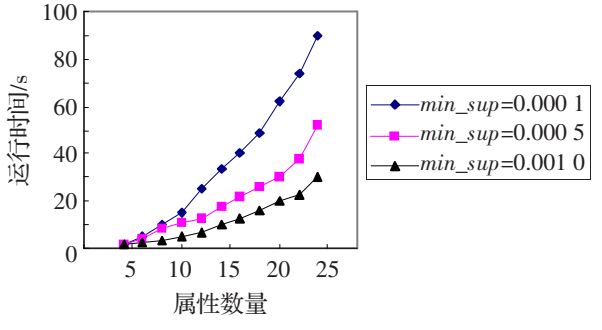


Fig.4　Cost by different attribute size

**图 4　属性数量和算法第三部分的运行时间变化关系**

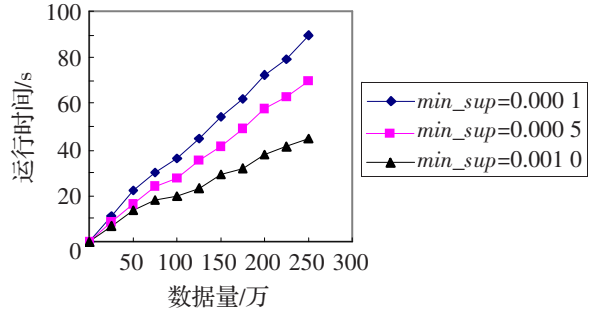


Fig.5　Cost by different data size

**图 5　数据量和算法第三部分的运行时间变化关系**

实验还使用了模拟数据来观察算法 2 对干预规则挖掘效率的提升效果。模拟数据量为 1 万条,属性数量为 100,每个属性有 3 个互斥项。关联规则的前

件全为 2 个项。图 6 显示了使用和不使用置信度记录 *ConfR* 的情况下算法 1 的不同效果。当初始关联规则数量增多时,需要考察的互斥项组合增加,减少对数据的扫描次数,有效地节约了时间。可见在规则数量较多的时候,使用算法 2 可以优化系统效率。

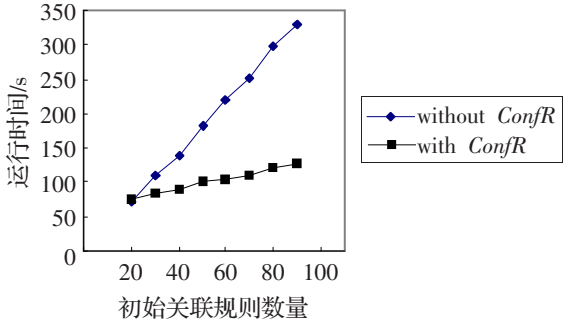


Fig.6 Cost by different number of AR

**图 6 关联规则数量和运行时间的变化关系**

## 5 结束语

本文提出了朴素干预规则的新概念,并将其应用于出生缺陷分析的实际问题中,取得了很好的效果。朴素干预规则还可以应用在一般的国家宏观调控决策、研究生扩招与就业决策等其他领域,朴素干预规则研究成果可以为决策者提供干预手段和干预强度的参考,有重要的社会意义。

## References:

[1] Wang Xufa, Chen Enhong, Wang Wei. Research on correlation of association rules[J]. Computer Engineering, 2000,26(7): 6-8.

[2] Agrawal R, Imielinski T, Swami A. Mining association rules between sets of items in large databases[C]//Proc 1993 ACM—SIGMOD Int Conf Management of Data. Washington IX: ACM Press, 1993:207-216.

[3] Li Hong, Cai Zhihua. Application of association rules in medical data analysis[J]. Microcomputer Development, 2003,13(6): 94-97.

[4] Ordonez C, Santana C A, de Braal L. Discovery interesting association rules in medical data[EB/OL]. [2000]. http://citeseer.nj.nec.com/ordonez00discovering.html.

[5] Brin S, Motwani R, Silverstein C. Beyond market: Generalizing association rules to correlations[C]//Processing of the ACM SIGMOD Conference 1997. New York: ACM Press, 1997: 265-276.

[6] Zhu Yuquan, Chen Geng, Yang Hebiao. Research on algorithm for mining positive and negative association rules[J]. Computer Science, 2006,33(3):188-190.

[7] Brin S, Motwani R, Silverstein C. Beyond market basket: Generalizing association rules to correlations[C]//Proc 1997 ACMSIGMOD Int'l Conf Management of Data, 1998:265-276.

[8] Agrawal R, Srikant R. Fast Algorithms for mining association rules in large databases[C]//20th Int'l Conf Very large databases, 1994:478-499.

[9] Han Jiawei, Pei Jian, Yin Yiwen. Mining frequent patterns without candidate generation[C]//Proc of 2000 ACM SIGMOD International Conference on Management of Data, Dallas, TX, 2000:1-12.

## 附中文参考文献:

[1] 王煦法,陈恩红,王玮.关联规则的相关性研究[J].计算机工程,2000,26(7):6-8.

[3] 李虹,蔡之华.关联规则在医疗数据分析中的应用[J].微机发展,2003,13(6):94-97.

[6] 朱玉全,陈耿,杨鹤标.正负关联规则挖掘算法研究[J].计算机科学,2006,33(3):188-190.

ZHANG Yue was born in 1983. She received the B.S. degree in Computer Science and Technology from Sichuan University in 2006. Now she is a M.S. candidate at Sichuan University. Her research interests include data mining, database and gene expression programming.

张悦(1983-),女,四川自贡人,2006 年于四川大学获计算机科学与技术专业学士学位,目前为四川大学计算机学院计算机应用专业硕士研究生,主要研究领域为数据挖掘,数据库,基因表达式编程。

TANG Changjie was born in 1946. He received the M.S. degree in Mathematics from Sichuan University in 1982. He is currently a full professor and doctoral supervisor at Sichuan University. His research interests include database and data mining.

唐常杰(1946-),男,重庆人,教授,博士生导师,1982年于四川大学数学系获得硕士学位,目前是四川大学计算机学院教授、博士生导师,主要研究领域为数据库,数据挖掘。

LI Chuan was born in 1977. He received the Ph.D. degree in Computer Science from Sichuan University in 2006. He is a lecturer at Sichuan University. His research interests include data mining, etc.

李川(1977-),男,四川成都人,2006 年于四川大学获得博士学位,目前任四川大学讲师,主要研究领域为数据挖掘等。

ZHU Jun was born in 1964. She received the M.S. degree in Embryo from Chongqing Medical University in 1988. She is a professor at National Center for Birth Defects Monitoring of Sichuan University. Her research interests include birth defect epidemiology and birth defect surveillance.

朱军(1964-),女,湖南长沙人,研究员,1988年于重庆医科大学获得硕士学位,目前是四川大学中国出生缺陷监测中心副主任/研究员,主要研究领域为出生缺陷流行病学和出生缺陷监测,发表论文近百篇,主编 3 本专著,主持国家"十一五"科技支撑计划课题等大型科研项目 20 多项。

ZENG Chunqiu was born in 1983. He is currently a M.S. candidate in Sichuan University. His research interests include data mining and gene expression programming.

曾春秋(1983-),男,四川安岳人,四川大学硕士研究生,主要研究方向为数据挖掘,基因表达式编程。

TANG Liang was born in 1983. He is currently a M.S. candidate. His research interests include data mining, database and knowledge engineering, etc.

唐良(1983-),男,重庆北碚区人,硕士研究生,主要研究方向为数据挖掘,数据库,知识工程。

LIU Xianbin was born in 1983. He is currently a M.S. candidate. His research interests include database and knowledge engineering, data mining.

刘显宾(1983-),男,江西瑞金人,硕士研究生,主要研究领域为数据库与知识工程,数据挖掘。