文章编号：2095-2163(2020)02-0001-07 中图分类号：TP391 文献标志码：A

面向健康医疗的分类关联规则挖掘研究

孙明瑞，臧天仪
（哈尔滨工业大学 计算机科学与技术学院，哈尔滨 150001）

摘 要：数据的爆炸式增长使知识和规则隐式地存在于数据中不被发现，形成资源孤岛，造成资源浪费。关联分析用以挖掘大规模、海量数据中隐式的关联规则模式。提出一种面向健康医疗的分类关联规则挖掘算法，用以挖掘大规模个人健康数据间隐藏的分类关联规则模式，对健康医疗及其它领域中基于关联规则的推荐具有新价值。

关键词：关联分析；频繁项集；分类关联规则；健康医疗

Research on classification association rule mining for health care

SUN Mingrui, ZANG Tianyi
(School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China)

【Abstract】 With the data increases exponentially, knowledge and rules are implicitly hidden in the data and are not discovered, forming resource islands and resulting in waste of resources. Association analysis is used to mine implicit association rules in large-scale, massive data. This paper proposes a classification association rule mining algorithm for health care, used to mine hidden classification association rules between large-scale personal health data, which could bring new value to association rule-based recommendations in health care and other areas.

【Key words】 association analysis; frequent item sets; classification association rules; health care

0 引 言

数据挖掘技术是从海量数据集中挖掘出用户感兴趣的物品或知识，这些知识是隐式的、未知的，挖掘出的知识表示为定律、规律、规则等形式。关联分析(Association Analysis)是数据挖掘的基础，用以挖掘大规模、海量数据中隐式的规则关系模式。大规模数据间存在关联关系，变量的取值也存在某种规律性，但数据间的关系是复杂的、隐式存在的，关联分析的目的就是挖掘数据间的隐藏关联信息，对健康医疗领域具有新的价值。典型的规则如购物篮事务(Market Basket Transaction)，“如果用户购买了婴儿尿布，那么该用户购买啤酒的概率为 33%”，产生的关系可以用关联规则(Association Rule)或频繁模式(Frequent Pattern)表示，用以反映用户偏好的有用规则。

基于以上的分析和讨论，本文提出了针对个人健康信息数据的频繁特征项挖掘算法，扩展了关联规则频繁模式的概念，引入连续性特征属性值并给出离散化的解决方案，改进了关联规则的一般模式，提出了分类关联规则挖掘算法，对健康医疗领域和其它领域中基于关联规则的推荐具有重要的指导意义和研究价值。

1 相关工作

关联规则挖掘是数据挖掘的一个重要研究方向，描述了交易数据集中 2 组不同对象之间存在的某种关联关系。在关联规则挖掘过程中，需要对交易数据集进行多次扫描并与候选频繁项目集进行匹配和计数。由于面对巨量交易数据集，这一匹配和计算过程需要花费大量时间，因此效率是设计算法的关键。

1994 年 Agrawal 等人[1]开创性地提出了 Apriori 算法，用来发现购物篮中有趣的关联关系，基于“所有长频繁项目集的子集都是频繁”的思想，对候选频繁项目集进行剪枝，使候选频繁项目集更小，从而显著改进频繁项目集算法的性能。自此，学界进行了广泛的研究，来解决关联分析中的实现和应用等问题。FP-growth[2]算法实现了 FP-tree 的构造及在 FP-tree 上进行挖掘，在效率上较之 Apriori 算法

基金项目：国家重点研发计划项目(2016YFC0901605, 2016YFC1201702-01)；国家高技术研究发展计划(2012AA02A601, 2015AA020101, 2015AA020108)。

作者简介：孙明瑞(1985-)，男，博士研究生，主要研究方向：服务计算、生物医学大数据计算、推荐算法等；臧天仪(1968-)，男，博士，教授，博士生导师，主要研究方向：服务计算与服务网络、生物医学大数据计算、数据密集型计算等。

通讯作者：孙明瑞 Email：mingrui.sun@gmail.com 臧天仪 Email：tianyi.zang@gmail.com

收稿日期：2019-06-10

有很大的提高。Park 等人[3]在 1995 年提出一种高效地产生频繁项目集的基于杂凑的 DHP 算法。Savasere 等人[4]设计了基于划分的算法。Toivonen[5]提出了一种基于采样的算法,其核心是随机从数据集中采集样本 S,然后搜索 S 中的频繁项集。惠晓滨等人[6]提出了基于频繁模式栈变换的高效关联规则算法。以上这些算法都是从算法实现问题的角度来解决频繁项集挖掘。

分类关联规则(Class Association Rules)是对关联规则的扩展,用以区分或判别实例类标签的关联规则。1998 年,Liu 等人[7]率先提出了分类关联规则算法 CBA。Wang 等人[8]融合分类关联规则和决策树的优势,提出关联决策树 ADT 算法。Xu 等人[9]首次提出利用原子型分类关联规则构建分类器的思想,创立了原子关联规则分类(CAAR)新技术。

综上所述,与传统的关联规则相比,分类关联规则具有更高的准确性及鲁棒性。作为一种新的特征匹配方法,如何处理连续属性特征,如何以较小的代价挖掘频繁项集,如何从大量关联规则中提取有效的分类关联规则,是本文研究的重点内容。

2 问题的定义

针对关联规则模式的挖掘,本次研究的目标是将数据的最后一列特征属性类别标识设置为关联规则的后件,即挖掘分类关联规则,而非全局关联规则。设 $I=\{i_1, i_2, \cdots, i_n\}$ 为具有 n 个特征属性的数据集合(项集约束),y 表示类别标识属性,且 $y \cap I = \varnothing$。分类关联规则的挖掘目标是形如 $X \rightarrow y$ 的分类关联关系,其中 $X \in I$。

3 算法研发与设计

3.1 先验性原理

一般地,包含 k 个项的数据集会产生 $2^k - 1$ 个频繁项集,由于 k 的值可能非常大,导致项集搜索空间的时间复杂度是指数级规模的 $O(2^n)$。因此引入如下定理:

定理 先验(Apriori)原理 如果一个项集是频繁的,则其所有的子集也一定是频繁的。

引理 如果项集是非频繁的,则其所有超集也都是非频繁的。

性质 单调性原理 设 I 是项的集合,$J=2^I$ 是 I 的幂集。度量 f 是单调的(向上封闭的),当

$$\forall X, Y \in J: (X \subseteq Y) \rightarrow f(X) \leqslant f(Y), \quad (1)$$

其中,X 为 Y 的子集,则 $f(X)$ 一定不会超过 $f(Y)$。

另一方面,f 是反单调的(向下封闭),当

$$\forall X, Y \in J: (X \subseteq Y) \rightarrow f(Y) \leqslant f(X). \quad (2)$$

根据性质,如图 1 所示,通常可以采用自顶向下(从 1 维到 n 维的项集搜索)或自底向上(从 n 维到 1 维的项集搜索)的搜索策略挖掘频繁项集。

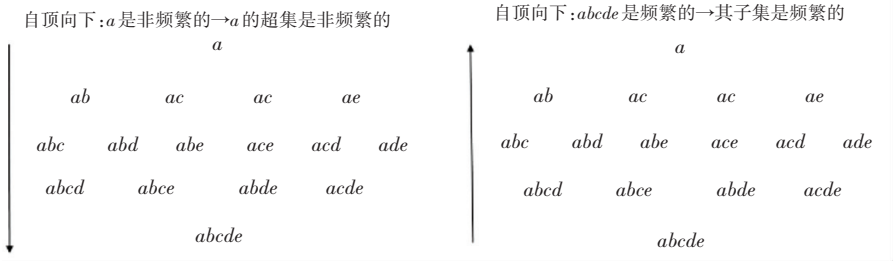

图 1 频繁项集封闭性原则

Fig. 1 Closure principle of frequent itemset

3.2 频繁项集挖掘算法

Apriori 算法基于支持度剪枝技术,采用分层的完备搜索算法(深度优先),依据性质向下封闭性,对关联规则进行挖掘,能有效防止项集的指数级增长。算法的功能是挖掘支持度大于等于 *minsup* 的项集。 算法首先生成单个元素项集列表,通过扫描数据集计算满足最小支持度的项集,删除不满足最小支持度的项集。对单个元素的项集进行组合以生成 2 个元素的项集。然后,重新扫描数据集,删除不满足最小支持度的项集,重复该过程直到所有项集都被删除。频繁项集挖掘算法的设计描述见如下。

算法 1 频繁项集挖掘算法

输入 数据集 I,最小支持度 *minsup*,频繁项集个数 $k_{\max}$

输出 频繁项集 F_k

$k \leftarrow$random integer from 1 to $k_{\max}$

If $k = 1$ then $F_k \leftarrow \{i \mid i \in I \wedge \sigma(i) \geqslant N \times minsup\}$ //发现所有的频繁 1-项集

Else do

$k \leftarrow k + 1$

$C_k \leftarrow apriori - gen(F_{k-1})$ //产生候选项集

For each $t \in T$ do

$C_t \leftarrow subset(C_k, t)$

//识别属于 t 的所有候选

For each $c \in C_t$ do

$\sigma(c) \leftarrow \sigma(c) + 1$

// 支持度计数增值

End for

End for

$F_k \leftarrow \{c \mid c \in C_k \wedge \sigma(c) \geqslant N \times minsup\}$　　// 提取频繁 k - 项集

Until $F_k = \varnothing$

Return $\cup F_k$

3.3　候选项集的挖掘和剪枝

候选项集挖掘通过合并频繁 $k-1$ 项集，当且仅当前 $k-2$ 项都相同。这里设 $X = \{x_1, x_2, \cdots, x_{k-1}\}$ 和 $Y = \{y_1, y_2, \cdots, y_{k-1}\}$ 是频繁 $k-1$ 项集，合并 X 和 Y，当其均满足如下公式：

$$x_i = y_i (i = 1, 2, \cdots, k-2) \text{ 且 } x_{i-1} \neq y_{i-1}. \quad (3)$$

3.4　算法的时间复杂度

对于频繁 1 项集的每个交易，都需要计算交易中每个项的支持度计数。设交易的平均宽度为 w，则 n 个交易的时间复杂度为 $O(nw)$。为了挖掘候选 k 项集，需要合并一对频繁 $k-1$ 项集，确定两者之间的 $k-2$ 个项相同，每次合并最多需要 $k-2$ 次相同项比较，最坏的情况下需要的总时间复杂度为 $O(\sum_{k=2}^{w}(k-2) \mid F_{k-1} \mid^2)$。

3.5　分类关联规则挖掘算法

针对分类关联规则挖掘，关联规则的后件为类别标识，即形如 $X \rightarrow y$ 的分类关联规则。通过 3.2 节频繁项集挖掘，研究得到了最大频繁项集 F_k，由其产生的子集可作为分类关联规则的前件，而目标类标识作为关联规则后件，是一种存在具体搜索目标的挖掘优化问题，同时需要满足设定的最小支持度和最小置信度阈值要求。这里，给出了分类关联规则挖掘算法的研发代码详见如下。

算法 2　分类关联规则挖掘算法

输入　频繁 k 项集 F_k

输出　分类关联规则 Car

For ($i = 2$; $F_{i-1} \neq \varnothing$; i ++) Do

$C_k \leftarrow Car - candidate - gen(F_{i-1})$;

For *each frequentitemset* $f \in F_k$ Do

For *each candidate* $c \in C_k$ Do

If $c.classset$ *is contained* in f Then

$c.classsupCount$ ++;

If $f.class = c.class$ Then

$c.rulesupCount$ ++;

End for

End for

$Car_k \leftarrow \{f \mid f \in F_k, f.rulesupCount / f.classsupCount \geqslant minconf\}$;

End for

Return $\cup Car_k$

4　实验评估

实验采用公开的临床数据集：加州大学欧文分校慢性肾病（Chronic Kidney Disease，CKD）数据集和皮肤病（Dermatology）数据集。考虑到数据集具有一定程度的缺失度，因此采用均值插值的方法对缺失值进行插值，并对连续属性项进行离散化分类处理。挖掘个人健康数据之间隐式关系、频繁项集、及分类关联规则模式。

慢性肾病数据集中有 400 个观测值，每个样本有 25 个属性项，其中 14 个是线性值类别属性项，11 个是连续数值属性项，目标类是二元指示器，表征患者是否患有慢性肾病（ckd 表示患有慢性肾病，notckd 表示不患有慢性肾病）。皮肤病数据集中有 366 个观测值，每个样本具有 35 个属性项。其中，34 个是线性类别属性项，1 个年龄属性是连续属性项，目标类为 6 种可能的疾病类型。

4.1　数据预处理

实验中，当对年龄区间进行离散化时，遇到的问题是如何确定区间的宽度。如果区间过宽，会因为缺乏置信度而丢失部分关联模式；如果区间过窄，会因为缺乏支持度而丢失部分关联模式；如果区间宽度定为 10 岁，将会导致某些置信度低于阈值，或者某些支持度低于阈值。因此，研究采用分位数离散化（Quantile Discretizer）将连续型数据转换成分类型数据。

对于连续属性项类别数目，研究中是依据已有特征属性项目的分类的最大值确定。因此采用 6 个类别来划分各属性值，详见表 1。例如，年龄属性（[2, 90]）通过字典序排序后，取用户数目的 6 个百分位数分成 6 个区间，即{(2, 34], (34, 46], (46, 54], (54, 60], (60, 67], (67, 90]}，确保转换后的年龄特征的类别平衡性。

实验使用的皮肤病数据集见表 2，年龄属性项采用 6 个类别进行离散化处理，即(0, 21], (21, 29], (29, 36], (36, 43], (43, 52], (52, 75]，确保转换后的年龄特征的类别平衡性。

表 1 实验所使用的慢性肾病数据集

Tab. 1 Chronic kidney disease data set used in the experiment

属性项	分类结果
连续属性项	
f_1. 年龄	(2,·34],·(34,·46],·(46,·54],·(54,·60],·(60,·67],·(67,·90]
f_2. 血压	(50,·60],·(60,·70],·(70,·76],·(76,·80],·(80,·90],·(90,·180]
f_{10}. 血糖随机	(22,·94],·(94,·108],·(108,·125],·(125,·148.2],·(148.2,·203],·(203,·490]
f_{11}. 血尿素	(1.5,·23],·(23,·32],·(32,·44],·(44,·53],·(53,·85],·(85,·391]
f_{12}. 血清肌酸酐	(0.4,·0.8],·(0.8,·1.1],·(1.1,·1.3],·(1.3,·2.2],·(2.2,·3.9],·(3.9,·76]
f_{13}. 钠	(4.5,·135],·(135,·137],·(137,·137.5],·(137.5,·139],·(139,·142],·(142,·163]
f_{14}. 钾	(2.5,·3.7],·(3.7,·4.2],·(4.2,·4.6],·(4.6,·4.62],·(4.62,·4.9],·(4.9,·47]
f_{15}. 血红蛋白	(3.1,·9.8],·(9.8,·11.4],·(11.4,·12.5],·(12.5,·13.7],·(13.7,·15.2],·(15.2,·17.8]
f_{16}. 血球容积比	(9,·31],·(31,·37],·(37,·38.9],·(38.9,·42],·(42,·47],·(47,·54]
f_{17}. 白血球数	(2 200,·6 300],·(6 300,·7 700],·(7 770,·8 406],·(8 406,·8 406.2],·(8 406.2,·9 800],·(9 800,·26 400]
f_{18}. 红细胞数	(2.1,·3.9],·(3.9,·4.7],·(4.7,·4.71],·(4.71,·4.8],·(4.8,·5.4],·(5.4,·8]
分类属性项	
f_3.:尿比重	1.005,·1.01,·1.015,·1.02,·1.025
f_4.:白蛋白	0,·1,·2,·3,·4,·5
f_5.:糖	0,·1,·2,·3,·4,·5|
f_6.:红细胞	abnormal,·normal
f_7.:Pus·cells 脓细胞	abnormal,·normal
f_8.:脓细胞团	notpresent,·present
f_9.:细胞域	notpresent,·present
f_{19}.:高血压	no,·yes
f_{20}.:糖尿病	no,·yes
f_{21}.:冠心病	no,·yes
f_{22}.:食欲	poor,·good
f_{23}.:踏板水肿	no,·yes
f_{24}.:贫血	no,·yes
f_{25}.:类别	nockd,·ckd

表 2 实验所使用的皮肤病数据集

Tab. 2 Dermatology data set used in the experiment

类别属性项	特征属性项	
	临床特征(值为 0,·1,·2,·3)	病理学特征(值为 0,·1,·2,·3)
C_1:·银屑病	f_1:·红疹	f_{12}:·黑色素失禁
C_2:·脂溢性皮炎	f_2:·鳞片排列	f_{13}:·嗜酸性粒细胞浸润
C_3:·扁平苔藓	f_3:·明确边界	f_{14}:·PNL 渗透
C_4:·玫瑰糠疹	f_4:·皮肤痒	f_{15}:·乳头状真皮纤维化
C_5:·慢性皮炎	f_5:·同形反应	f_{16}:·胞外分泌
C_6:·毛发红糠疹	f_6:·多边形丘疹	f_{17}:·棘皮症
	f_7:·滤泡性丘疹	f_{18}:·眼角膜细胞增多
	f_8:·涉及口腔粘膜	f_{19}:·角化不全
	f_9:·涉及膝关节及肘关节	f_{20}:·棒状突起的形成
	f_{10}:·涉及头皮	f_{21}:·棱脊的延伸
	f_{11}:·家族史(0 *or* 1)·	f_{22}:·上表皮变薄
		f_{23}:·海绵状脓疱
		f_{24}:·芒罗微小脓肿
		f_{25}:·焦状颗粒层增厚
		f_{26}:·颗粒层的消失
		f_{27}:·基底层液泡化及破坏
		f_{28}:·海绵层水肿
		f_{29}:·视网膜锯齿状外观
		f_{30}:·滤泡性角插头
		f_{31}:·毛囊周角化不全
		f_{32}:·炎症性单核细胞浸润
		f_{33}:·带状渗透
		f_{34}:·年龄·(linear)…(0,·21),·(21,·29],·(29,·36],·(36,·43],·(43,·52],·(52,·75]

4.2　CKD 数据集频繁项集及关联规则挖掘

研究设定数据集中最后一列为目标类(ckd=患病,notckd=未患病),也就是分类关联规则的后件,实验的最小支持度 $minsup = 0.15$,最小置信度 $minconf = 0.6$,挖掘出最大的频繁 $k = 13$ 项集数量为 3 个,由此产生的关联规则模式见表 3(截取前 6 个强关联规则模式)。

表 3　CKD 数据集关联规则模式

Tab. 3　Association rule patterns for chronic kidney disease data sets

分类关联规则模式	支持度	置信度
高血压=yes=> ***class = ckd***	0.367 5	1
糖尿病=yes · ==> · ***class = ckd***	0.342 5	1
细菌域=notpresent · 高血压=yes · ==> · ***class = ckd***	0.337 5	1
脓细胞团=notpresent · 高血压=yes · ==> · ***class = ckd***	0.300 0	1
高血压=yes · 冠心病=no · ==> · ***class = ckd***	0.292 5	1
脓细胞团=notpresent · 糖尿病=yes · ==> · ***class = ckd***	0.282 5	1

由表 3 可知,当用户患有高血压疾病时,患有慢性肾病的支持度为 36.75%,置信度为 1;当用户患有糖尿病时,患有慢性肾病的支持度为 34.25%,置信度为 1;当这两种疾病与其它个人健康数据状况并发发生时,支持度略有下降。关联规则模式中另外一条记录(高血压=yes 糖尿病=yes ==> class=ckd sup:0.265 $conf$:1),表示同时患有高血压和糖尿病的患者罹患慢性肾病的样例共有 106 例,置信度为 1。此实验的研究结果与慢性肾病权威机构 NIDDK 提出的影响因素完全吻合[10]。因此,如果一名患者满足上述条件,那么该患者就有极大的可能性发展成为慢性肾病,对个人健康的影响风险巨大。针对此情况,患者要做好预防慢性肾病的准备。通过对分类关联规则模式和规律的分析,就可以挖掘有意义的信息来评价患者的身体状况。通过研究结果,最终揭示慢性肾病致病因素与个人健康数据状况的关联关系,要控制并发症的产生,防止疾病进一步恶化的情况出现,使患者能够及时接受治疗,为用户健康提供科学的服务推荐方法。

4.3　皮肤病数据集频繁项集及关联规则挖掘

研究中,设定数据集中最后一列为目标类(银屑病=1,皮脂性皮炎=2,扁平苔藓=3,玫瑰糠疹=4,慢性皮炎=5,毛发红糠疹=6),即关联规则的后件,本文中实验的最小支持度 $minsup = 0.15$,最小置信度 $minconf = 0.6$,挖掘出最大的频繁 $k = 14$ 项集数量为 1 个,由此产生的关联规则模式见表 4。截取 6 个强关联规则模式,其中 2 个分类关联规则模式的目标类为“皮脂性皮炎=2”和“扁平苔藓=3”。

表 4　皮肤病数据集关联规则模式

Tab. 4　Association rule patterns for dermatology data sets

分类关联规则模式	支持度	置信度
多边形丘疹=0 · 乳头状真皮纤维化=0 · 胞吐=0 · 海绵水肿=0 · 滤泡性角插头=0==> · ***class*** · = 1	0.254 1	1
多边形丘疹=0 · 乳头状真皮纤维化=0 · 胞吐=0 · 海绵水肿=0 · 毛囊周角化不全=0 · ==> · ***class*** · = 1	0.254 1	1
口腔粘膜介入=0 · 乳头状真皮纤维化=0 · 胞吐=0 · 海绵水肿=0 · 滤泡性角插头=0 · ==> · ***class*** · = 1	0.254 1	1
口腔粘膜介入=0 · 乳头状真皮纤维化=0 · 胞吐=0 · 海绵水肿=0 · 毛囊周角化不全=0 · ==> · ***class*** · = 1	0.254 1	1
同形反应=0 · 滤泡性丘疹=0 黑素失禁=0 · 乳头状真皮纤维化=0 雷特山脊的棒状突起=0 乳头上表皮变薄=0 颗粒层消失=0 · ==> · ***class*** · = 2	0.166 7	0.95
滤泡性丘疹=0 膝关节和肘关节受累=0 · PNL 渗透=0 · 乳头状真皮纤维化=0 雷特山脊的延伸=0 颗粒层消失=0 · ==> · ***class*** · = 3	0.273 2	0.61

由实验可知,银屑病($class = 1$)的关联规则模式数量远大于其它类型皮肤病关联规则模式的数量。以表 4 中第一条数据为例,若某个皮肤病患者的临床症状满足“多边形丘疹=0,乳头状真皮纤维化=0,胞吐=0,海绵水肿=0,滤泡性角插头=0”,那么该患者的皮肤病类型为银屑病的支持度为 25.41%,置信度为 1。因此,如果一名皮肤病患者满足上述条件,那么就有极大的可能性是患有银屑病类型的皮肤病,医生会根据银屑病的临床症状与疾病特点,对患者对症下药。通过对皮肤病关联规则模式和规律的分析,可以挖掘出有意义的数据信息来对患者所属的皮肤病类别加以区分,针对疾病的特点进行治疗,为患者提供个性化的服务推荐方法。

4.4　算法性能评估

本节将讨论算法的性能。首先研究在不同最小支持度和最小置信度下的算法运行时间,如图 2 所示。2 个数据集实验结果显示,最小置信度的变化对运行时间的性能影响不大,但是当最小支持度变小时运行时间会呈指数级增长。

对于构建的关联规则数量,如图 3 所示,关联规则的数量与运行时间的情况类似,都是随着最小支持度的降低,而呈现指数级的增长。

频繁项集的数量,只与最小支持度有关,与最小置信度无关,如图 4 所示。频繁项集的数量也是随

着最小支持度的降低成指数级增长。这些实验结果表明,最小支持度,也就是项集的反单调性,对剔除非频繁项集是非常有效的。

图 2 不同最小支持度和最小置信度的运行时间

Fig. 2 Running time with different minimum support and minimum confidence

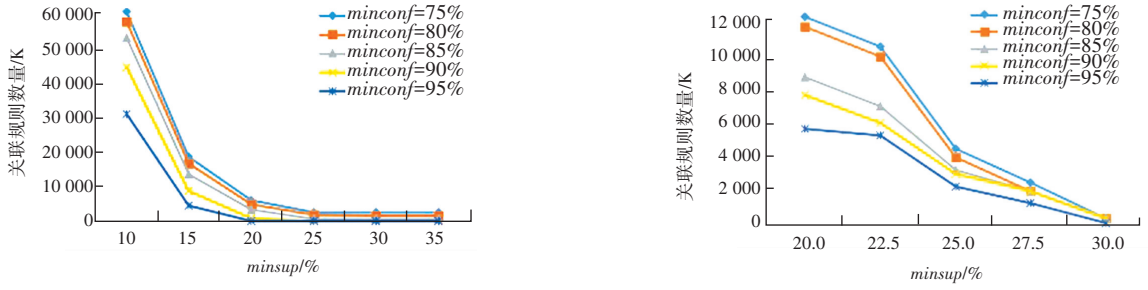

图 3 不同最小支持度和最小置信度的关联规则数量

Fig. 3 Number of association rules with different minimum support and minimum confidence

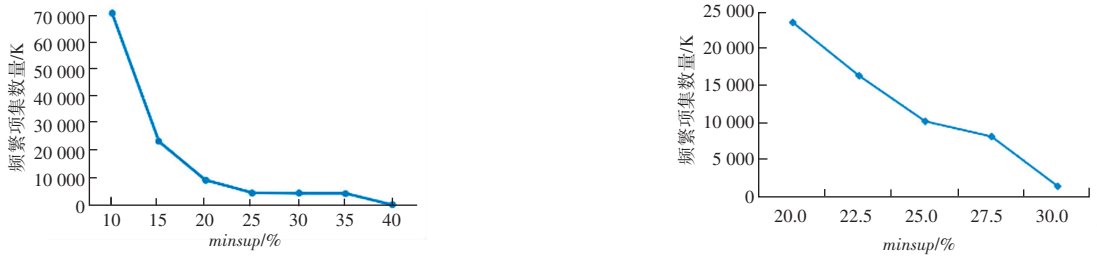

图 4 不同最小支持度的频繁项集数量

Fig. 4 Number of frequent itemsets with different minimum supports

综上所述,如果最小支持度 *minsup* 较高,比如达到 30%,此时就可以用较短的运行时间获得较少而有意义的关联规则,当然其中的很多规则可能是平凡的。如果为了挖掘更有意义的关联规则模式,可以采用较小的最小支持度,但这会导致无法接受的系统运行时间和指数级关联规则和频繁项集。因此采用合理的、可接受的最小支持度和最小置信度的值,是非常必要的。

5 结束语

本文提出从数据的隐式特征中识别出数据间的分类关联规则模式,通过支持度和置信度的权衡保证分类关联规则模式的有效性。基于关联分析的算法,对数据内在特征的分类关联规则模式进行挖掘,实现了个人健康数据的分解及分类,挖掘出数据的频繁项集,发现致病因素的重要数据特征,对个人健康特征的疾病预防与推荐具有重要的指导和建设性意义。

参考文献

[1] AGRAWAL R, SRIKANT R. Fast algorithms for mining association rules[C]// Proceedings of the 20th VLDB Conference. Santiago: Morgan Kaufmann, 1994: 487.

[2] HAN J W, PEI J, YIN Y W. Mining frequent patterns without candidate generation[C]// ACM SIGMOD´2000. Dallas, TX: ACM, 2000: 1.

[3] PARK J S, CHEN M, YU P S. An effective hash-based algorithm for mining association rules [C]//Proceeding of the ACM SIGMOD International Conference on Management of Data. New York:ACM, 1995: 175.

[4] SAVASERE A, OMIECINSKI E, NAVATHE S B. An efficient algorithm for mining association rules in large databases[C]// VLDB´1995, Proceedings of 21th International Conference on Very Large Data Bases. San Francisco, CA: Morgan Kaufmann Publishers Inc., 1995:432.

(下转第 11 页)