

基于概念格的多值属性关联规则可视化

郭晓波^{1,2,3}, 赵书良^{1,2,3*}, 赵娇娇^{1,2,3}, 刘军丹^{1,2,3}

(1. 河北师范大学 数学与信息科学学院, 石家庄 050024; 2. 河北省计算数学与应用重点实验室(河北师范大学), 石家庄 050024;

3. 河北师范大学 移动物联网研究院, 石家庄 050024)

(* 通信作者电子邮箱 zhaoshuliang@sina.com)

摘要:针对传统关联规则可视化方法无法展现数据间的频繁模式和关联关系,表示形式比较单一,缺乏多模式展现形式等问题,提出了一种新的多值属性关联规则可视化表示算法。该算法运用概念格理论对多值属性数据进行重新定义和分类,将频繁项集和关联规则中的多值数据项分别以概念格结构进行表示,实现了频繁项集可视化展示和一对一、一对多、多对一、多对多及概念分层的多模式关联规则可视化展示。最后,以某省全员人口数据为基础对算法进行了具体实现和分析,同时实现了对人口数据的源数据、频繁模式以及关联关系的可视化展示。实验结果表明,所提出的可视化形式和已有成果相比具有良好的频繁项集与多模式关联规则展现效果。

关键词:多值属性;概念格;关联规则;可视化;人口数据

中图分类号: TP311 **文献标志码:** A

Visualization of multi-valued attribute association rules based on concept lattice

GUO Xiaobo^{1,2,3}, ZHAO Shuliang^{1,2,3*}, ZHAO Jiaojiao^{1,2,3}, LIU Jundan^{1,2,3}

(1. College of Mathematics and Information Science, Hebei Normal University, Shijiazhuang Hebei 050024, China;

2. Hebei Key Laboratory of Computational Mathematics and Applications (Hebei Normal University), Shijiazhuang Hebei 050024, China;

3. Institute of Mobile Internet of Things, Hebei Normal University, Shijiazhuang Hebei 050024, China)

Abstract: Considering the problems caused by the traditional association rules visualization approaches, including being unable to display the frequent pattern and relationships of items, unitary express, especially being not conducive to represent multi-schema association rules, a new visualizing algorithm for multi-valued association rules mining was proposed. It introduced the redefinition and classification of multi-valued attribute data by using conceptual lattice and presented the multi-valued attribute items of frequent itemset and association rules with concept lattice structure. This methodology was able to achieve frequent itemset visualization and multi-schema visualization of association rules, including the type of one to one, one to many, many to one, many to many and concept hierarchy. At last, the advantages of these new methods were illustrated with the help of experimental data obtained from demographic data of a province, and the source data visualization, frequent pattern and association relation visual representation of the demographic data were also achieved. The practical application analysis and experimental results prove that the schema has more excellent visual effects for frequent itemset display and authentic multi-schema association rules visualization.

Key words: multi-valued attribute; concept lattice; association rule; visualization; demographic data

0 引言

在数据挖掘研究领域中,关联规则(association rules)是一个重要的研究方向,其作用是从数据集中发现属性间存在的、隐藏的、新颖的、有趣的关联或相关关系,从海量数据中获取信息和知识。然而,一般方法却无法将数据间存在的频繁模式和关联模式以可视化的形式展现出来,不能帮助用户获取更为完备的信息。作为知识的一种可视化表示形式,概念格(concept lattice)已经被人们应用到很多研究领域。概念格将哲学的概念进行数学化的描述,实现了概念的一种形式化描述,其表达数据的基本形式是形式背景。在大量数据库应用中,对于数据的分析并非都是单值属性的形式背景——单值背景^[1],更多的是复杂多值属性的形式背景——多值背景^[2-3]。Bal等^[4]给出了基于形式背景分析的频繁项集搜索

与关联规则提取的可视化方法,但该方法无法处理多值属性数据。Cassio等^[5]采用着色和变形技术从概念格提取多值数据并对其进行树形可视化展示,该模式能够表示数据项之间的概念关系,不足之处是展现形式灵活性较差,不支持用户交互性操作,用户无法动态分析数据之间频繁模式和关联关系。Julien等^[6]利用可视化后处理方法进行交互式关联规则挖掘,主要对一对一形式的关联规则进行展示,但无法展示一对多、多对一和多对多形式的关联规则。Michael等^[7]介绍了关联规则的分层展示形式,该形式不利于用户对挖掘结果进行多层次关联分析,并且展示结果容易出现部分重叠现象。Dario等^[8]对8类关联规则的可视化展现方法进行了综合分析,这些方法一般适用于布尔类型数据,而不利干处理多值属性数据,无法满足用户分析与展现多值属性项之间关系的需求。

收稿日期:2013-02-21;修回日期:2013-04-04。

基金项目:河北省科学技术研究与发展计划项目(072435158D, 09213515D, 09213575D);河北师范大学硕士基金资助项目(201102002)。

作者简介:郭晓波(1986-),男,河北栾城人,硕士研究生,CCF会员,主要研究方向:数据挖掘、智能信息处理; 赵书良(1967-),男,河北献县人,教授,博士生导师,博士,主要研究方向:数据挖掘、智能信息处理; 赵娇娇(1986-),女,河北清苑人,硕士研究生,主要研究方向:自然语言处理、智能信息处理; 刘军丹(1987-),女,河北临城人,硕士研究生,主要研究方向:应用数学、智能信息处理。

目前,多数关联规则可视化研究工作主要集中于挖掘结果的可视化展示,大都存在以下不足:所采用的方法不利于展现多值属性数据的频繁模式与关联关系^[9]、缺少挖掘过程的交互性与可视化^[10-11]、用户无法动态分析规则信息^[12]。最重要的是关联规则表示形式比较单一,无法对频繁项集进行可视化展示及关联规则多模式展现,用户难以动态地分析数据项之间的频繁模式和关联模式。

本文提出一种新的基于概念格的多值属性关联规则可视化方法,结合概念格理论对多值属性数据进行了重新定义和分类,给出了频繁项集和多值属性关联规则可视化表示算法。通过引入概念格结构把数据项有机地组织起来,使数据之间的关系通过概念格节点的特化关系与例化关系生动简洁地表达出来,不仅便于用户对频繁项集进行可视化展示和动态分析,而且实现了一对一、一对多、多对一、多对多以及概念分层的多模式关联规则可视化展示。利用概念格理论提出了多值属性关联规则可视化的完整解决方案,通过数据源可视化、可视化数据挖掘过程及交互式参数调整、挖掘结果的可视化等机制,使用户可代替领域专家直接进行数据挖掘,大幅提高了规则的展现效果和挖掘结果的可用性。

1 多值属性关联规则的概念格表示

1.1 项目集的概念格表示

在实际应用中,全员人口数据库的育龄妇女人口记录通常以形式背景(formal context)表示对象集的基本形式,为了更好地将事务集以概念格的形式进行表示,这里将项集与概念格相结合,研究概念格与频繁项目集之间的关系。

定义1 属性又称为项目,设 $A = \{a_1, a_2, \dots, a_n\}$, a_k ($k \in \mathbf{N}^+, 1 \leq k \leq n$, a_k 称为一个项目,表示一个属性) 为 n 个不同项目的集合。设定事务集 $T = \{(t_1, i_1), (t_2, i_2), \dots, (t_k, i_k), \dots, (t_m, i_m)\}$ ($k \in \mathbf{N}^+, 1 \leq k \leq m$), 其中 (t_k, i_k) 代表一个概念,表示一个对象(事务), t_k 是对象(事务)的标识符, $i_k \subseteq A$ 是对象(事务)的属性值(项目)。

给定一个三元组 (T, I, R) 称为形式背景,其中 T 是事务的有限集合, I 是属性值的有限集合, R 是 $T \times I$ 上的二元关系,存在唯一的偏序集合与之相对应,并且由这种偏序集合产生一种格结构,这种由 (T, I, R) 所诱导的格 L 称为一个概念格^[1]。

概念格中的每个节点 $N = (T_k, I_k)$ 是一个二元组,其中 $T_k \in P(T) = 2^T, I_k \in P(I) = 2^I, P(T)$ 和 $P(I)$ 分别表示数据库中的事务集和项目集的幂集,定义如下映射:

$$f: 2^T \rightarrow 2^I, f(T_k) = T_k' = \{i \mid i \in I, \forall t \in T_k \subset T, tRi\}$$

$$g: 2^I \rightarrow 2^T, g(I_k) = I_k' = \{t \mid t \in T, \forall i \in I_k \subset I, tRi\}$$

并且 $T_k' = I_k, I_k' = T_k$, 则 T_k 和 I_k 分别称为概念的外延和内涵,其中 f 和 g 称为 T 的幂集和 I 的幂集之间的 Galois 连接^[1]。

定义2 设 $H = \{1, 2, \dots, k\}$ 是描述数据项概念层的集合,对于两个概念 $(i_1, k_1), (i_2, k_2)$, 其中 $i_1, i_2 \in T \times H, k_j$ 表示数据项 i_j 所属概念层,若 $i_1 \subset i_2$, 则 $k_1 < k_2, k_1, k_2 \in H$ 。

定义3 设定事务集 $T = \{(t_1, i_1), (t_2, i_2), \dots, (t_m, i_m)\}$ 是由一系列形式背景组成的集合,对于任意 (t_i, i_i) 和 (t_j, i_j) 具有相同的层关系 k ($k \in \mathbf{N}^+$), 则存在 $(t_1, i_1), (t_2, i_2), \dots, (t_k, i_k) \in T$ 使得 $(t_i, i_i) = (t_1, i_1) > < (t_2, i_2) > < \dots > < (t_k, i_k) = (t_j, i_j)$, 其中 $> <$ 表示具有相同的层关系, T 可以表示为有序集合 $(T; > <)$ 。概念层 $T_k(t, i)$ 是一组概念 (t_i, i_i) 集合,其中每个 (t_i, i_i) 具有相同的层关系 $k, k \in H$ 。

1.2 多值属性数据分类

所谓多值背景^[16]就是事务(记录)和属性之间不能仅仅用布尔型关系来表示,而是在原有的形式背景中出现了属性值的集合,并用具体的属性值来表示。比如,在某省全员人口数据库中,“学历”、“文化程度”、“年龄”、“户口性质”等均称为多值背景,即事务与属性之间的关系无法只用“1”或“0”表示。为了便于挖掘任务的实现,本文提出适合多值属性关联规则可视化挖掘的多值背景定义,根据属性的类别分为三类,具体介绍如下。

在多值属性集中,对于“年龄”“世代间隔^[17]”等这样的表示数量化的属性项,其属性值都是用具体的数值来描述事务与属性之间的关系,则称该多值背景为数值型多值背景,其定义如下:

定义4 设五元组 (T, I, N, H_N, R_N) 是一个数值型多值背景。其中: T 是事务集, I 是属性集, N 是数值型属性值的集合, H_N 是数值属性的概念层集合,而 $R_N \subseteq T \times I \times N \times H$ 是它们之间存在的一个四元关系。当且仅当对于任意 $t \in T, i \in I, h \in H$, 有且只有一个 $n \in N$ 满足 $(t, i, n, h) \in R_N$, 用 $(t, i, n, h) \in R_N$ 表示“对于属性 i , 事务 t 在 h 上具有数值型属性 n ”。若满足 $(t, i, n_j, h_j) \in R_N$ 且 $(t, i, n_{j'}, h_{j'}) \in R_N$, 那么必有 $n_j = n_{j'}, h_j = h_{j'}$, 其中 $j \in \mathbf{N}^+$, 表明 T 中同一 I 的数值性属性值的集合 N 在 H_N 上相等。

在实际应用中,很多属性项所具有的属性值为区间形式,即属性值都是以具体的区间值来描述事务与属性之间的关系,则称该多值背景为区间型多值背景,其定义如下:

定义5 设五元组 (T, I, S, H_S, R_S) 是一个区间型多值背景,其中: T 是事务集, I 是属性集, S 是区间属性值的集合, H_S 是区间型的概念层集合,而 $R_S \subseteq T \times I \times S \times H$ 是表示它们之间存在的一个四元关系。当且仅当对于任意 $t \in T, i \in I, h \in H$, 有且只有一个 $s \in S$ 满足 $(t, i, s, h) \in R_S$, 用 $(t, i, s, h) \in R_S$ 表示“对于属性 i , 事务 t 在 h 具有区间型属性 s ”。若满足 $(t, i, s_j, h_j) \in R_S$ 且 $(t, i, s_{j'}, h_{j'}) \in R_S$ 那必有 $s_j = s_{j'}, h_j = h_{j'}$, 其中 $j \in \mathbf{N}^+$ 。即 $s_j^L = s_{j'}^L, s_j^U = s_{j'}^U, h_j = h_{j'}$, 其中: $s_j \in [s_j^L, s_j^U], s_{j'} \in [s_{j'}^L, s_{j'}^U]$ (s_j^L, s_j^U 表示 s_j 的最小、最大取值, $s_{j'}^L, s_{j'}^U$ 表示 $s_{j'}$ 的最小、最大取值), 表明 T 中同一个 I 的 S 在 H_S 上相等。

对于多值属性集中,如文化程度分为“高级”、“中级”和“初级”等,其属性值都是以具体的类别值来描述事务与属性之间的关系,则称该多值背景为类别型多值背景,其定义如下:

定义6 设五元组 (T, I, C, H_C, R_C) 是一个类别型多值背景,其中: T 是事务集, I 是属性集, C 是类别型属性值的集合, H_C 是类别型的概念层集合,而 $R_C \subseteq T \times I \times C \times H$ 是它们之间存在的一个四元关系。当且仅当对于任意 $t \in T, i \in I, h \in H$, 有且只有一个 $c \in C$ 满足 $(t, i, c, h) \in R_C$, 用 $(t, i, c, h) \in R_C$ 表示“对于属性 i , 事务 t 在 h 上具有类别型属性 c ”。若满足 $(t, i, c_j, h_j) \in R_C$ 且 $(t, i, c_{j'}, h_{j'}) \in R_C$, 那么必有 $c_j = c_{j'}, h_j = h_{j'}$, 其中 $j \in \mathbf{N}^+$, 表明 T 中同一个 I 的 C 在 H_C 上相等。

1.3 多值属性关联规则表示

对于任意 $a \in I, a$ 的取值可以为数值型、区间型和类别型。设 a 的取值集合为 V , 若满足 $\forall v \in V_n$ 存在 $v, \mu \in \mathbf{N}^+, v \leq \mu$ 使得 $v \in [v, \mu]$, 则称 a_n 为数值型多值属性; 若满足 $\forall v \in V_s$ 存在 $l, v \in \mathbf{N}^+$ 使得 $v = [l, v]$, 则称 a_s 为区间型多值属性; 若满足 $\forall v \in V_c = [\alpha_1, \alpha_2, \dots, \alpha_m]$ ($m \in \mathbf{N}^+$), 则称 a_c 为类别型多值属性。

如果 a_n 为数值型属性值, $a_n = \langle a, v, \mu \rangle$ (其中 $\langle a, v, \mu \rangle \in I \times \mathbf{N}^+ \times \mathbf{N}^+$), 则三元组 $\langle a, v, \mu \rangle$ 表示数值属性 a 的属性值在区间 $[v, \mu]$ 上; 如果 a_s 为区间型属性值, $a_s = \langle a, l,$

v) (其中 $\langle a, l, v \rangle \in I \times N^+ \times N^+$), 则三元组 $\langle a, l, v \rangle$ 表示数值属性 a 的属性值是 $[l, v]$; 如果 a_c 为类别型属性值, $a_c = \langle a, \alpha \rangle$ (其中 $\langle a, \alpha \rangle \in I \times N^+$), 则二元组 $\langle a, \alpha \rangle$ 表示属性 a 的属性值为 α 。由此可知, 类别属性只与值相关, 而数值或区间属性既可以与值相关联, 也可以与区间相关联。元组 $\langle a, v, \mu \rangle$ 、 $\langle a, l, u \rangle$ 和 $\langle a, \alpha \rangle$ 称为项 (Item), 则 I 称为项集 (ItemSet), $\langle i \rangle$ 是项集 i 所包含的属性集合。

定义 7 若对于任意 $\langle a, v, \mu \rangle$ 、 $\langle a, l, u \rangle$ 和 $\langle a, \alpha \rangle \in \langle i \rangle$, 存在 $\langle a, q \rangle \in t_i$, 使得 $v \leq q \leq \mu$ 或 $q = [l, v]$ 或 $q = \alpha$ 成立, 则称事务 t_i 支持 i 。

定义 8 多值属性关联规则是具有 $i_l \Rightarrow i_r$ 形式的蕴涵式, 其中 $i_l, i_r \subset I$, 并且 $\langle i_l \rangle \cap \langle i_r \rangle = \emptyset$ 。如果 T 中有 $s\%$ 的事务支持 i_l 和 i_r , 且 $c\%$ 的支持 i_l 的事务也支持 i_r , 则该规则的支持度和置信度为 $s\%$ 和 $c\%$ 。

2 算法描述

传统的关联规则可视化方法不利于展现多值属性数据的频繁模式与关联关系, 无法实现可视化展示频繁项集与多模式展现关联规则, 用户难以动态地分析数据项之间的频繁模式和关联模式。针对这些问题, 本文引入关键属性参数 (Key Attribute Factor, KAF) 和概念层参数 (Concept Hierarchy Factor, CHF) 进行多值属性关联规则展示, 方便用户有选择性地展示和分析, 较大地提升了频繁项集和关系规则的展示效果。

为了提高展现模式的交互性和用户友好性, 频繁项集可视化算法采用概念格结构将频繁项集有机地组织起来, 使得数据之间的关系通过概念格节点的特化关系与例化关系进行体现。

算法 1 频繁项集可视化算法 VFreqItems()。

输入: 频繁项集 L , 最小支持度 $minSup$, KAF 参数, CHF 参数;
输出: 频繁项集可视化形式。

VFreqItems (L , $minSup$, KAF , CHF)

- 1) $freqItems = get_key_freqItem (L, KAF, CHF)$;
//根据 KAF 和 CHF 参数选择频繁项集
- 2) FOR $item_i \in freqItems$ && $item_i.sup \geq minSup$ DO {
- 3) 根据每个频繁项集 $item_i$ 中所包含的项数将其划分到相应的层次 L_k 上; // $k = item_i.count$
- 4) 每层 L_k 上的节点以支持度大小递增排序, 调用 Line (L_k , N_{ki}) 将同层节点画到一条直线上;
// L_k 表示第 k 层的节点集合, N_{ki} 表示第 k 层第 i 个节点
- 5) FOREACH $N_{ki} \in L_k$ DO {
- 6) FOREACH $N_{(k+1)j} \in L_{k+1}$ DO {
- 7) IF $N_{ki} \subset N_{(k+1)j}$ THEN
- 8) $VF = VF \cup \{N_{ki} \rightarrow N_{(k+1)j}\}$;
// N_{ki} 指向其父节点 $N_{(k+1)j}$
- 9) } //end foreach
- 10) } //end foreach
- 11) } //end for
- 12) return VF;

在 $KAF\{N_i, S_i, C_i\}$ 中: N_i 表示数值型属性值集合, S_i 表示区间型属性值集合, C_i 表示类别型属性的集合, 如 $KAF\{(\text{年龄}), (\text{间隔}), (\text{户口性质})\}$ 。而在 $CHF\{N_j, S_j, C_j\}$ 中: N_j 表示数值型属性值的概念层数, S_j 表示区间型属性值的概念层数, C_j 表示类别型属性的概念层数, 如 $CHF\{3, 3, 3\}$ 。首先, 算法 1 利用 $minSup$ 、KAF 参数和 CHF 参数选择满足条件的频繁项集, 遍历每个频繁项集, 并调用 Line (L_k, N_{ki}) 函数将所有的项集依据其所包含的项数目和支持度大小进行分层布局; 然

后, 遍历每层上的每个节点, 将当前节点指向其所有父节点, 并减少与其他节点的重叠, 保证布局层次清晰。最后, 输出频繁项集的可视化形式。VFreqItems() 的优点是: 由于引进关键属性参数和概念层参数来定义查询的数据集, 这使得产生冗余项集的问题在该类算法中也得到了很好的解决。可视化结果布局结构良好, 层次清晰, 便于动态分析数据项之间的频繁模式, 克服了同类算法中缺少频繁项集可视化展示的不足。Line (L_k, N_{ki}) 函数用来减少节点边与边之间的交叉, 保证同层节点在一条直线上, 并将所有节点按支持度大小进行排序, 有利于提升频繁项集可视化效果。

算法 1 的执行实例如图 1 所示。第一步依据 $minSup$ 、KAF 参数和 CHF 参数大小选择频繁项集 $F01 \sim F11$, 扫描频繁项集表。第二步将项集节点 $F01 \sim F11$ 划分到不同的层上, 如图 1 所示将所有项集节点分为四层显示。第三步将每层中项集节点 F_i 以支持度 Sup 的大小进行排序; 如第一层 $\{F01, 1\}$, $\{F03, 1\}$, $\{F04, 1\}$, $\{F02, 1\}$, $\{F05, 1\}$ (后面的数字代表节点所属层数)。第四步根据每个 F_i 节点的 FaP 值, 将每层中项集节点指向其父节点; 如 $F02$ 的父节点指针 $FaP_{F02}\{6, 8\}$, 连接形式 $\{F02, 1\} \rightarrow \{F06, 2\}$, $\{F02, 1\} \rightarrow \{F08, 2\}$ 。最后生成频繁项集可视化形式。

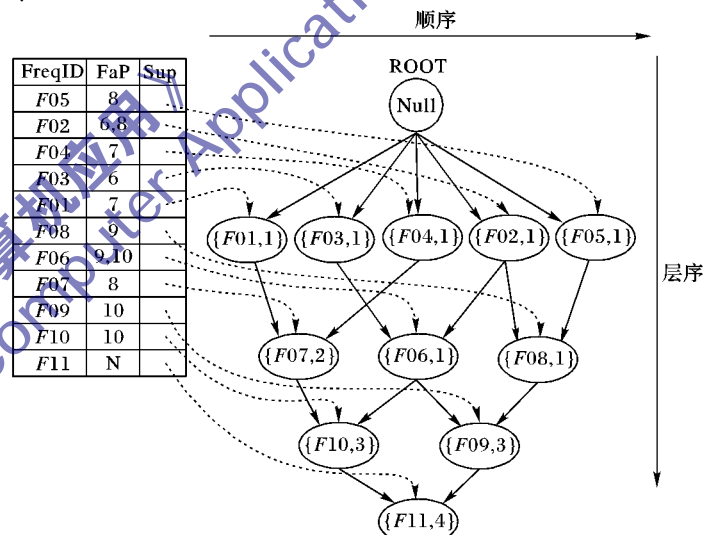


图1 频繁项集可视化

多值属性关联规则可视化算法采用概念格结构对关联规则进行可视化展示, 通过设置 RM (Representation Modal) 和 CLN (Cross Level Number) 值的大小来构建不同模式关联规则, 帮助用户动态分析规则信息, 满足不同用户需求, 其中 RM 表示关联规则可视化的展示形式: 默认形式 (一对一、一对多、多对一、多对多) 和概念分层形式; CLN 表示规则前件与后件所允许跨层分析的层数。

算法 2 关联规则可视化算法 VRules()。

输入: 关联规则 $RuleSet$, 最小置信度 $minConf$, CLN 跨层数, RM 展现模式;

输出: 关联规则可视化形式。

VRules ($RuleSet$, $minConf$, CLN , RM)

- 1) 根据 CLN 的值选择规则可视化模式;
//规则前件与后件所允许跨层分析的层数
- 2) IF $RM = \text{"默认"}$ THEN {
- 3) 根据每个规则前件 R_{lhs} 和后件 R_{rhs} 所包含项的个数, 将其划分到相应的层次 L_k 上; // $k = R_{lhs.count}$ 或 $R_{rhs.count}$
- 4) ELSE {
- 5) 根据每个规则前件 R_{lhs} 和后件 R_{rhs} 中项的概念层 $item_i$ CLN 值, 将其划分到相应的层次 L_k 上; // $k = R_{lhs.chf}$ 或 $R_{rhs.chf}$


```

6)      } //end if
7)  FOR  $R \in RuleSet$  &&  $R.conf \geq minConf$  DO { //R 表示规则
8)      调用  $Round(L_k, N_{ki})$  确定各层节点的具体位置  $N_{ki}$ ,
          使同层节点在同一个圆周上;
          //  $L_k$  表示第  $k$  层的节点集合,  $N_{ki}$  表示第  $L$  层第  $i$  个节点
9)      FOREACH  $N_{ki} \in L_k$  &&  $N_{ki} \in R_{lhs}$  DO {
          // 获取  $k$  层的规则前件节点  $N_{ki}$ 
10)     FOREACH  $N_{(k+CLN)j} \in L_{k+CLN}$  &&
           $N_{(k+CLN)j} \in R_{rhs}$  DO {
          // 取第  $(k + CLN)$  层规则后件节点  $N_{(k+CLN)j}$ 
11)     IF  $N_{ki} \in R$  &&  $N_{(k+CLN)j} \in R$  THEN
12)          $VR = VR \cup \{N_{ki} \rightarrow N_{(k+CLN)j}\}$ ;
13)     } //end foreach
14) } //end foreach
15) } //end for
16) return VR;

```

在算法2描述中,首先,根据 RM 和 CLN 的值选择可视化模式;然后,遍历所有规则,调用 $Round(L_k, N_{ki})$ 函数依据规则前件和后件中所包含的项的个数(或项的概念层)将其前后件划分到不同层的圆周上;其次,遍历每个规则的前件,将其指向规则后件并生成一条关联规则;最后,输出相应的关联规则可视化结果。 $VRules()$ 的主要优点是:实现了多模式关联规则可视化展示形式,允许用户灵活选择各种形式的关联规则进行分析和研究,解决了同类可视化算法中关联规则表示形式比较单一,无法进行同层、跨层、不同概念层间的规则分析和挖掘的问题。为了提高规则的友好性和展示效果,通过 $Round(L_k, N_{ki})$ 函数确定各层节点的具体位置,避免节点重叠,使同层节点分布在同一个圆周上,使得展示形式具有3D的形式。

算法2的执行实例如图2所示。第一步按 $CLN = 1$ (表示规则前件节点与后件节点在同层或相隔一层)和 $RM =$ “默认”选择关联规则可视化形式,扫描关联规则表选择满足大于 $minConf$ 的关联规则。第二步将关联规则中的前、后件节点 $L01 \sim L12$ 划分到不同层的圆周上,如图2所示将规则节点分为三层显示,如第二层 $\{04, 2\}$, $\{05, 2\}$, $\{06, 2\}$ 和 $\{07, 2\}$ 。第三步根据每个规则 R 前件节点的规则后件(Right Hand Side, RHS)值,将每层中规则前件节点指向其对应的规则后件节点。如前件节点 $L04 \{5, 7, 8\}$, 生成规则 $R: \{04, 2\} \rightarrow \{05, 2\}$, $R: \{04, 2\} \rightarrow \{07, 2\}$, $R: \{04, 2\} \rightarrow \{08, 3\}$ (第二项表示项数)。最后生成关联规则可视化结果。

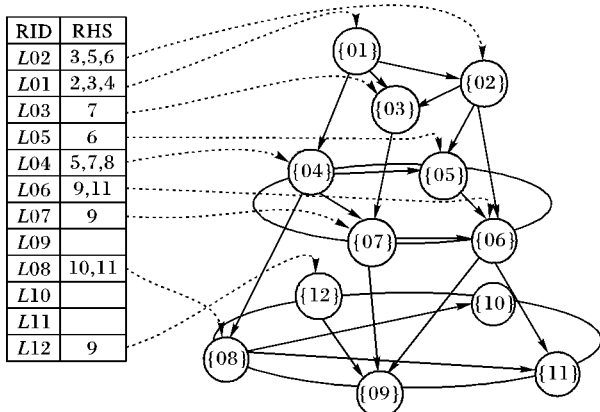


图2 关联规则可视化

3 关联规则可视化挖掘过程

3.1 源数据可视化

源数据可视化阶段,运用可视化技术将数据库中的数据

以“第二语言”——图形的形式进行展示,通过选择数据集和多种数据展示工具帮助用户进行专业的数据分析,以便挑选针对性和关联性更强的数据来进行分析和研究,使其不再局限于通过关系表来分析数据信息,而且能够以更直观的方式观察数据及其结构关系。

3.2 频繁项集可视化

频繁项集可视化挖掘阶段运用概念格结构展示频繁项集,具有表示形式清晰、挖掘过程灵活、用户交互性强等优点,使得频繁项集更容易被人们理解。该阶段主要包括:点支持度($MinSup$)或区间($[MinSup, MaxSup]$)查询,频繁项集个数($K-Item$)或区间($[K_{min}, K_{max}]$)查询,动态调整 KAF 参数与 CHF 参数,指定频繁查询模式,分层挖掘,上卷、下钻和附属信息分析功能。例如:将支持度区间设置为 $[78, 92]$, 查询支持度在区间内的所有频繁项集,让用户从量上对挖掘出来的频繁项集进行科学分析,从中发现有价值的信息。

3.3 关联规则可视化

关联规则可视化挖掘阶段,主要采用概念格结构对关联规则进行展示。通过设置 $minConf$ 、 KAF 参数和 CHF 参数,构建不同模式关联规则,形成多模式可视化展示。同时,允许用户对同层间、跨层间、不同概念层间的规则进行分析和挖掘,极大地满足了用户的不同需要。该阶段功能包括:附属信息显示,点置信度($MinConf$)或区间($[MinConf, MaxConf]$)查询, KAF 和 CHF 参数调整,设置规则前件(LHS)和后件(RHS)的个数及包含项,一对一、一对多、多对一、多对多和概念分层关联规则展示形式。

某省全员人口库中的人口记录包含大量的多值属性字段,例如:文化程度、户口性质、人口所属地区和育龄妇女世代间隔等,这些数据项或属性所隐含的概念具有层次关系,在低层或原始抽象层的数据项之间很难找出强关联规则,而在较高的抽象层发现的强关联规则可能提供更具有价值的信息。本文将概念分层纳入到关联分析中,采用离散化方法对数据库的多值数据进行处理,方便用户分析不同概念层或跨层间的数据关系。如某省全员人口数据库中:文化程度{初级{小学,初中},中级{高中,大专}};地区{盆地{柳江,宣化},高原{沽源,康保}}。用户可以对其进行概念分层形式的可视化展示,并针对不同概念层的数据进行关联性分析,从中挖掘有用的信息,制定科学合理的决策。

4 关联规则可视化应用实例

本文以某省全员人口数据库为数据源,对源数据、频繁项集和规则可视化进行了具体实现。下面是对育龄妇女世代间隔的大小与育龄妇女的文化程度、年龄、所属地区和户口性质之间的频繁模式和关联关系进行了具体分析。

4.1 育龄妇女数据的源数据可视化

人口数量的增减是由子女一代人数与父母一代人数的比例决定的。当子女一代人数与父母一代人数相等,形成一个静止人口时,平均生育年龄愈低,两代人时间间隔愈短,在平均预期寿命相同的情况下,同时存在的人口数目就会愈多^[13]。针对某省全员人口数据的特点,从库中分别选取人口所属地区为山地、平原、丘陵、盆地和高原的育龄妇女信息,对妇女世代间隔进行分析。首先,以年龄树的形式展示女性年龄分布情况,从图3中可得到育龄妇女人口数量,运用正态分布函数对所选择记录的世代间隔进行分析,得到这些地区的育龄妇女世代间隔集中分布在22~24,如图4所示,以便对其进行深入分析和研究。

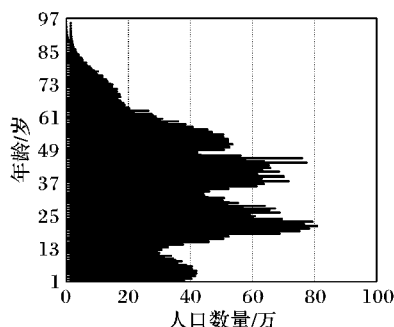


图3 女性人口年龄金字塔

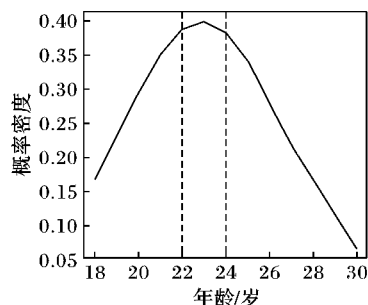


图4 育龄妇女世代间隔分布

4.2 育龄妇女数据的频繁项集可视化

对数据库记录进行离散化处理,调用算法1对频繁项集进行可视化展示。育龄妇女世代间隔的大小与育龄妇女的文化程度、年龄、所属地区和户口性质有很大关系,利用算法1来展示它们之间的频繁模式,通过设置 $KAF\{(\text{年龄}), (\text{间隔}), (\text{文化}, \text{地区}, \text{户口性质})\}$, $CHF\{1,1,1\}$ 参数和 $minSup = 35$ 从全员人口数据抽取文化程度、年龄、所属地区、户口性质和世代间隔字段进行分析,从全员库中挖掘出隐含的频繁模式,以概念格结构进行可视化展示,如图5所示。

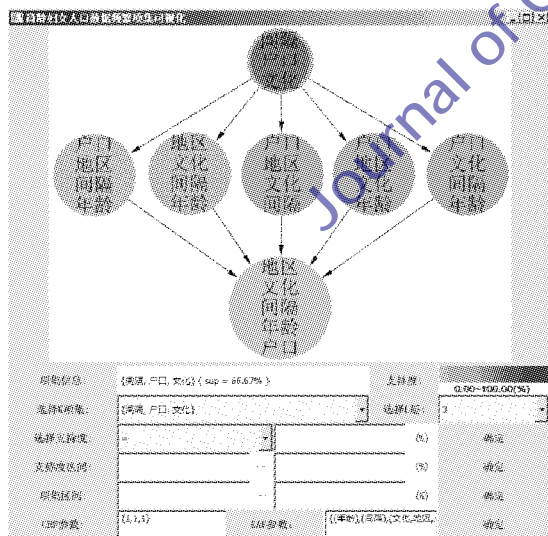


图5 频繁项集可视化功能界面

如图5所示,用颜色的深度来表示频繁项集支持度的大小,将每层频繁项集的支持度按照从小到大的方式进行排序,方便用户观察和分析感兴趣的频繁项集;同样可设定支持度的大小来选择相关项;采用概念格的形式来展示频繁项集,结构层次鲜明清晰。最为重要的是,通过设置其他参数挖掘用户所需要的频繁项集,其中包括选择支持度区间、指定相关频繁项等,并且可以对频繁项集进行上卷和下钻操作(如图6所示),极大地提高了挖掘过程的可操作性,提升了用户体验。

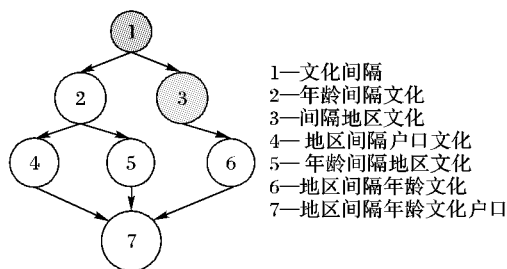


图6 选择指定频繁项集可视化展示实例

4.3 育龄妇女数据的关联规则可视化

鉴于大多数关联规则挖掘形式无法进行多模式展示,为了满足用户的多样性需求,本文提出的关联规则可视化展示形式能够有效地展现多模式关联规则,对4.2节所展示的频繁项进行多模式关联规则可视化表示,调用关联规则可视化算法2,设置 $KAF\{(\text{年龄}), (\text{间隔}), (\text{文化}, \text{地区}, \text{户口性质})\}$, $CHF\{2,2,2\}$ 和 $minConf \geq 50$ 来对数据库中包含文化程度、年龄、所属地区、户口性质和世代间隔的记录集进行可视化展示,如图7所示。

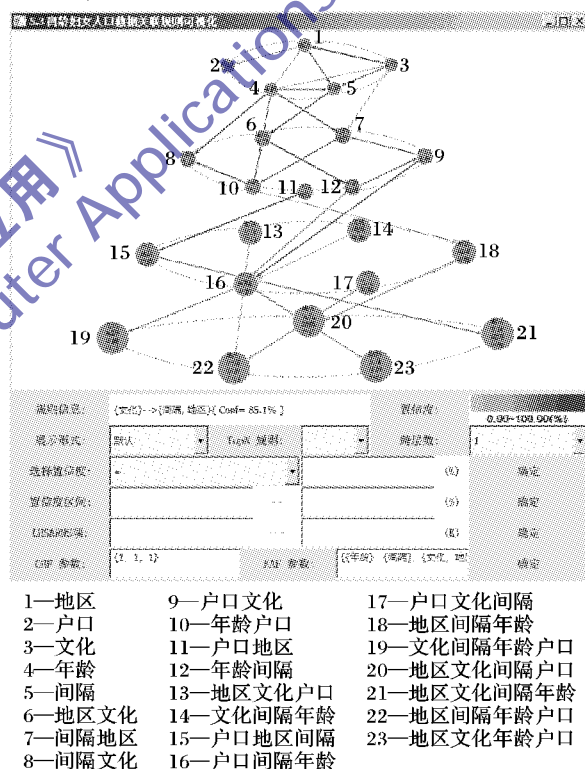


图7 关联规则可视化展示功能界面

由图8可以看到,运用概念格形式来展示关联规则具有良好的特性,其展示结果具有较强的立体效果和友好的表达方式。同4.2节,用规则前后件连接线的颜色和粗细来表示不同规则对应的置信度的大小,提高用户的视觉感知能力。

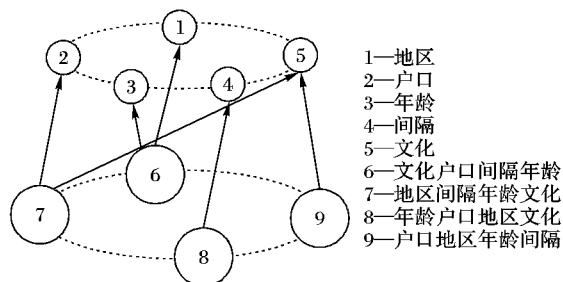


图8 关联规则可视化展示实例

相关介绍如下:首先,在 LHS&RHS 框中输入“5”“1”显

示多对一模式,其中 $LHS = 5$ 和 $RHS = 1$ 表示规则前后件中包含的属性值个数,如图9所示。

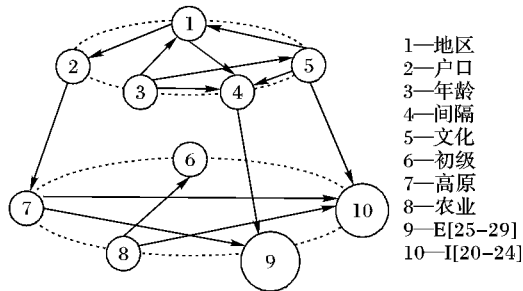


图9 多对一展示形式

其次,同样可以在 LHS&RHS 框中输入“3”“2”来分析多对多模式规则,如图10所示,以直观、清晰的形式展示规则前件和后件中包含多项的关联规则,极大地满足了用户的不同需求,实现多模式关联规则可视化。

同需求,实现多模式关联规则可视化。

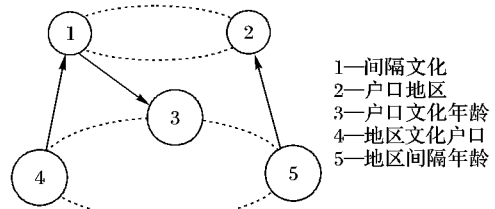


图10 多对多展示形式

然后,通过设置 $KAF\{(年龄), (间隔), (地区, 文化)\}$ 、 $CHF\{3, 3, 3\}$ 、 $CLN = 1$ 和 $minConf \geq 18$ 以概念分层的形式展示年龄、间隔、地区和文化之间不同抽象层之间的关系,如图11所示,用户分析不同概念层的规则,不仅能从宏观上对整个关联规则进行把握,而且可以从微观层面对规则进行更为详细的分析和研究,得出更深入的、更有说服力的信息。

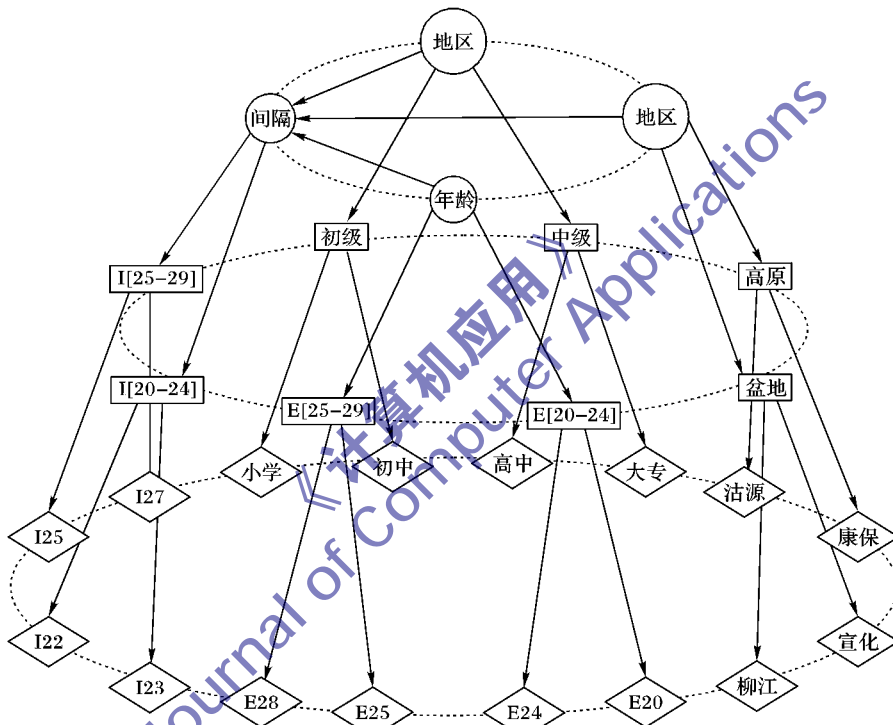


图11 概念分层展示形式

本文实现的多值属性关联规则可视化形式与基于表、二维矩阵、TwoKey 图、Double-Decker 图以及平行坐标的规则可视化^[8-9,14-15]相比具有以下优点:

实现了对多值属性数据的频繁项集和关联规则挖掘,方便用户动态分析不同类型字段项之间的关系和频繁模式;实现了频繁项集可视化展示,可对挖掘结果进行上卷、下钻以及分层查询操作,另外支持 K -Item、支持度区间查询和附加信息展示;实现了一对一、一对多、多对一、多对多和概念分层的多模式关联规则可视化,且不对所展示规则信息的前件与后件的项的数量进行限制,可查询多对多对模式的规则;规则前后件信息区分明显,可解释性较好;规则之间不易重叠,避免出现界面混乱的现象;用户可灵活选择各种感兴趣的规则展示模式,完成不同层次间的规则分析。

5 结语

通过对多值属性数据的分析与研究,本文提出一种新的基于概念格的多值属性关联规则可视化方法,实现了对多值

属性数据的频繁项集可视化展示与一对一、一对多、多对一、多对多和概念分层的多模式关联规则可视化展示,便于用户动态分析多值属性数据之间的频繁模式和相关关系。通过运用某省全员人口数据对算法进行了具体实现和分析,实验结果表明本文所提出的关联规则可视化表现形式具有良好的显示效果和用户交互性,在很大程度上提高了用户体验,实现了多值属性关联规则可视化挖掘。

在下一步的工作中,将研究如何利用频繁项集和关联规则中所含数据项之间的语义联系与应用背景,把频繁项集和规则转换为领域知识进行可视化知识展示。

参考文献:

- [1] GANTER B, WILLE R. Formal concept analysis: mathematical foundations [M]. Berlin: Springer-Verlag, 1999: 17-35.
- [2] GUGISCH R. Many-valued context analysis using descriptions [C]// Conceptual Structures: Broadening the Base, LNCS 2120. Berlin: Springer-Verlag, 2001: 157-168.

(下转第2211页)

4、5所示。海量数据下的相似数据检测在查全率和时间效率是相互矛盾的。从表4中可以看出,相同的运行时间下,RGM的查全率略低于IWM,从表5中可以明显地看到,在不同数据量上,IWM的运行速度比RGM快。这都是因为IWM采用多线程并行检测技术、加速法和优先队列技术,大大减少记录比对时间和总体测时间,既保证查全率又减少检测时间;而RGM为了保证查全率,采用多趟检测技术,增加了检测时间。

表4 两种算法相同运行时间条件下的查全率比较

数据量(万)	同时间下查全率/%	
	IWM	RGM
53.4	98.2	97.6
98.1	97.3	96.6
126.2	95.9	95.3
153.7	95.0	94.5

表5 两种算法相同查全率条件下消耗时间对比

数据量(万)	同查全率下消耗时间/min	
	IWM	RGM
53.4	8.6	22.1
98.1	14.8	29.8
126.2	21.7	37.8
153.7	28.3	46.4

综上所述,本文提出的基于海量数据的相似重复记录检测算法的性能要优于基于等级分组的相似检测算法。

4 结语

针对海量数据下相似重复记录检测问题,本文采取了多种有效策略。首先采用主观因素和客观因素综合考虑的综合加权法计算各属性的权重,然后采用多线程依据各属性对数据集并行排序,使用加速法提前结束记录比对算法;最后合并检测结果集。实验结果表明,该方法是一个合理、有效的相似重复数据检测方法。本文方法仍有许多未解决的问题,例

如:记录之间的相似度阈值大小是根据经验设定的。由于它对记录的检测精度有一定的影响,所以将在以后的工作中继续研究阈值的设定问题。

参考文献:

- [1] MONGE A E, ELKAN C P. The field matching problem: algorithms and applications [C]// Proceedings of the 2nd Conference on Knowledge Discovery and Data Mining. Cambridge: AAAI, 1996: 267-270.
- [2] MINTON S N, NANJO C, KNOBLOCK C A, et al. A heterogeneous field matching method for record linkage [C]// Proceeding of the 5th IEEE International Conference on Data Mining. Piscataway: IEEE, 2005: 314-321.
- [3] HERNANDEZ M, STOLFO S. The merge/purge problem for large databases [C]// Proceedings of the 1995 ACM SIGMOD International Conference on Management of Data. New York: ACM, 1995: 127-138.
- [4] BLENK O M, MOONEY R. Adaptive name matching in information integration [J]. IEEE Intelligent Systems, 2003, 18(5): 16-23.
- [5] 邱越峰,田增平,季文赟,等.一种高效的检测相似重复记录的方法[J].计算机学报,2001,24(1):69-77.
- [6] 鲁均云,李星毅,施化吉,等.基于内码序值聚类的相似重复记录检测方法[J].计算机应用研究,2010,27(3):874-878.
- [7] 孟祥逢,鲁汉榕,郭玲,等.基于遗传神经网络的相似重复记录检测方法研究[J].计算机工程与设计,2010,31(7):1550-1553.
- [8] 李星毅,包从剑,施化吉.数据仓库中的相似重复记录检测方法[J].电子科技大学学报,2007,36(6):1273-1277.
- [9] MONGE A E, ELKAN C. An efficient domain-independent algorithm for detecting approximately duplicate database records [C]// Proceedings of the SIGMOD 1997 Workshop on Research Issues on Data Mining and Knowledge Discovery. Cambridge: AAAI, 1997: 23-29.
- [10] 张永,迟忠先.位置编码在数据仓库ETL中的应用[J].计算机工程,2007,33(1):50-52.
- [11] NGUYEN T T, C HUI S C, CHANG K Y. A lattice-based approach for mathematical search using formal concept analysis [J]. Expert Systems with Applications, 2012, 39(5):5820-5828.
- [12] BAL M, BAL Y, USTUNDAG A. Knowledge representation and discovery using formal concept analysis: an HRM application [C]// WCE 2011: Proceedings of the World Congress on Engineering. London: Newswood, 2011:1068-1073.
- [13] CASSIO M, LEGRAND B. Extracting and visualising tree-like structures from concept lattices [C]// IV'11: Proceedings of the 2011 15th International Conference on Information Visualisation. Washington, DC: IEEE Computer Society, 2011:261-266.
- [14] JULIEN B, FABRICE G, HENRI B. Interactive visual exploration of association rules with rule-focusing methodology [J]. Knowledge and Information Systems, 2007, 13(1):43-75.
- [15] MICHAEL H, CHELLUBOINA S. Visualizing association rules in hierarchical groups [C]// Interface 2011: Statistical, Machine Learning, and Visualization Algorithms. Cary, North Carolina: SAS Institute, 2011:1-11.
- [16] DARIO B, CRISTINE D. Visual mining of association rules [C]// Visual Data Mining: Theory, Techniques and Tools for Visual Analytics, LNAI 6208. Berlin: Springer-Verlag, 2008:103-122.
- [17] BILAL A, ERHAN A, ALI K. MODENAR: Multi-objective differential evolution algorithm for mining numeric association rules [J]. Applied Soft Computing, 2008, 8(1):646-656.
- [18] PACHON A V, VAZQUEZ J. An evolutionary algorithm to discover quantitative association rules from huge databases without the need for an a priori discretization [J]. Expert Systems with Applications, 2012, 39(1):585-593.
- [19] MARTINEZ B M, RIQUELME J. Analysis of measures of quantitative association rules [C]// HAIS'11: Proceedings of the 6th International Conference on Hybrid Artificial Intelligent Systems. Berlin: Springer-Verlag, 2011:319-326.
- [20] SHAHARANEE M, HADZIC F, DILLON S. Interestingness measures for association rules based on statistical validity [J]. Knowledge-Based Systems, 2011, 24(3):386-392.
- [21] SAMUEL Y, MEKITIE W, MULUMEBET A, et al. Duration and determinants of birth interval among women of child bearing age in Southern Ethiopia [J]. BMC Pregnancy and Childbirth, 2011, 11(38):1-6.
- [22] SONG S J, KIM E H, KIM H E, et al. Query-based association rule mining supporting user perspective [J]. Computing, 2011, 93(1):1-25.
- [23] LIU G M, ANDRE S. AssocExplorer: an association rule visualization system for exploratory data analysis [C]// KDD '12: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2012:1536-1539.
- [24] GELYB A, RAOUL M, NOURINE L. Representing lattices using many-valued relations [J]. Information Sciences, 2009, 179(16):2729-2739.
- [25] 马瀛通.人口统计分析学[M].北京:红旗出版社,1989:696.

(上接第2203页)