一种新的面向普通用户的多值属性关联规则可视化挖掘方法

郭晓波[1,2,3],赵书良[1,2,3],王长宾[1],陈 敏[1,2,3]
(1.河北师范大学数学与信息科学学院,河北石家庄 050024;2.河北省计算数学与应用重点实验室,河北石家庄 050024;
3.河北师范大学移动物联网研究院,河北石家庄 050024)

摘 要: 针对传统关联规则可视化挖掘方法不利于处理多值属性数据、缺乏展现数据间的频繁模式和关联模式以及效率低下等问题,提出了基于 KAF 因子和 CHF 因子的 Apriori 改进算法进行多值属性关联规则挖掘,实现了一种新的基于概念格的多值属性关联规则可视化方法.运用概念格理论对多值属性数据进行了重新定义和分类,建立了较为完整的挖掘过程参数调整策略,方便用户选择关键属性值进行规则挖掘分析,提高了算法运行速度和挖掘效率.以概念格结构将多值数据组织起来,实现了对频繁项集的可视化展示,以及关联规则的多模式可视化展示.实验结果表明,改进后的挖掘算法具有更好的性能,所提出的可视化形式和已有成果相比具有良好的展现效果.

关键词: 多值属性; 概念格; 关联规则; 可视化挖掘
中图分类号: TP391.13 **文献标识码:** A **文章编号:** 0372-2112 (2015)02-0344-09
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2015.02.021

A New Visualizing Mining Method of Multi-Valued Attribute Association Rules for Ordinary Users

GUO Xiao-bo[1,2,3], ZHAO Shu-liang[1,2,3], WANG Chang-bin[1], CHEN Min[1,2,3]
(1. *Mathematics & Information Science College, Hebei Normal University. Shijiazhuang, Hebei* 050024, *China*;
2. *Hebei Key Laboratory of Computational Mathematics and Applications. Shijiazhuang, Hebei* 050024, *China*;
3. *Institute of Mobile Internet of Things, Hebei Normal University. Shijiazhuang, Hebei* 050024, *China*)

Abstract: Considering the problems aroused by the traditional association rules visualization mining methods which are lack of dealing with multi-valued attribute data, especially not conducive to expressing the frequent pattern between items and representing multi-schema association rules, this paper, which presents the redefinition and classification of multi-valued attribute data by using conceptual lattice, proposes an improvement of Apriori algorithm based on the KAF factor and the CHF factor to mine multi-valued attribute association rules as well as introduces a novel visualizing approach for multi-valued association rules based on concept lattice, and establishes a complete mining course parameters adjustment strategy acting very well in improving the speed and efficiency of mining algorithm, which is convenient for users to select key attribute values to mine and analyze rules. This methodology organically organizes the multi-valued attribute data with concept lattice structure, which has achieved frequent itemset visualization and multi-schema visualization of association rules. The experimental results turn out that the improved mining algorithm has a better performance and the schema has much excellent visual effects for multi-schema association rules visualization.

Key words: multi-value attribute; concept lattice; association rules; visualizing mining

1 引言

在数据挖掘研究领域中,关联规则可视化挖掘是一个重要的研究方向,其目标是借助可视化技术从数据集中发现属性间隐藏的价值信息.然而,现有一些方法不能有效地处理海量数据集中多值属性数据、无法简洁直观地展现出数据间存在的频繁模式和关联模式.作为数据分析的有力工具,概念格已经被人们应用到数据挖掘研究中,诸如基于量化概念格的关联规则挖掘分析方法[1]、基于格和哈希表的关联规则挖掘方法[2]、自适应

收稿日期:2013-12-09;修回日期:2014-02-21;责任编辑:蓝红杰
基金项目:国家自然科学基金(No.71271067);国家社会科学基金(No.13BTY011)

方法[3]和多维关联规则挖掘[4]等算法相继出现.另外还有一些借助人类视觉能力实现规则可视化挖掘: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[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[6]、利用可视化后处理方法进行交互式关联规则挖掘[7]、关联规则的分层展示[8].Dario 等[9]对 8 类关联规则的可视化展现技术进行了综合分析,这些方法一般适用于布尔类型数据,而不利于处理多值属性数据,无法满足用户分析与展现多值属性项之间频繁模式和关联关系的需求.

目前,海内外研究人员对关联规则可视化挖掘的研究内容主要集中于对挖掘结果的可视化展示,大都存在以下不足[10~14]:不利于处理多值属性数据;缺乏有效的挖掘参数调整策略;用户无法挑选针对性较强的数据进行关联规则挖掘.最重要的是关联规则表示形式相对比较单一,无法实现对频繁项集及关联规则多模式展现.

为了更好地处理海量数据中的多值属性的形式背景—多值背景[15,16],本文结合概念格理论对多值属性数据进行了重新定义和分类,建立完整的参数调整策略,将关键属性因子 KAF(Key Attribute Factor)和概念层因子 CHF(Concept Hierarchy Factor)引入到 Apriori 算法中进行多值属性关联规则挖掘;结合挖掘算法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,以概念格结构将多值数据组织起来,实现了对频繁项集的可视化展示与一对一、一对多、多对一、多对多和概念分层的关联规则可视化展示,便于用户对频繁项集及关联规则进行动态地分析和研究.

2　多值属性关联规则的概念格表示

2.1　项目集的概念格表示

定义 1　属性又称为项,设 $A=\{a_1,a_2,\cdots,a_k,\cdots,a_n\}$(其中,$k\in N^+$,$1\leqslant k\leqslant n$,$a_k$ 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 $T=\{(t_1,i_1),(t_2,i_2),\cdots,(t_k,i_k),\cdots,(t_m,i_m)\}$(其中,$k\in N^+$,$1\leqslant k\leqslant m$),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)集合.

给定一个三元组(T,I,R)称为形式背景(Formal Context),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,存在惟一的偏序集合与之对应,并且由这种偏序集合产生一种格结构,这种由形式背景 $T,I,R)$所诱导的格 L 称为一个概念格[15].

定义 2　设 $H=\{1,2,\cdots,k,\cdots,maxk\}$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 $T=\{(t_1,i_1),(t_2,i_2),\cdots,(t_m,i_m)\}$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 $k(k\in N^+)$,则存在$(t_1,i_1),(t_2,i_2),\cdots,(t_k,i_k)\in T$使得$(t_i,i_i)=(t_1,i_1)><(t_2,i_2)><\cdots><(t_k,i_k)=(t_j,i_j)$,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,$k\in H$.

2.2　多值属性数据分类

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,$R_N\subseteq T\times I\times N\times H$ 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 $j\in N^+$,表明 T 中同一个 I 的 N 在 H_N 上相等.

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 $j\in N^+$.即 $s_j^L=s_j^{L'}$,$s_j^U=s_j^{U'}$,$h_j=h_j'$,其中 $s_j=[s_j^L,s_j^U]$,$s_j'=[s_j^{L'},s_j^{U'}]$,表明 T 中同一个 I 的 S 在 H_S 上相等.

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 $j\in N^+$,表明 T 中同一个 I 的 C 在 H_C 上相等.

2.3　多值属性关联规则表示与求解

对于任意 $a\in I$,a 的取值可以为数值型、区间型和类别型.

定义 7　设 a 的取值集合为 V,若满足 $\forall v\in V_n$ 存在 $\upsilon,\mu\in N^+$,$\upsilon\leqslant\mu$ 使得 $v\in[\upsilon,\mu]$,则称 a_n 为数值型多

值属性.如果 a_n 为数值型属性值,$a_n = <a,\upsilon,\mu>$(其中 $<a,\upsilon,\mu> \in I \times N^+ \times N^+$),则三元组 $<a,\upsilon,\mu>$ 表示数值属性 a 的属性值在区间$[\upsilon,\mu]$上.

定义 8 若满足 $\forall v \in V_s$ 存在 $l,v \in N^+$ 使得 $v = [l,v]$,则称 a_s 为区间属性.如果 a_s 为区间型属性值 $a_s = <a,l,v>$(其中 $<a,l,v> \in I \times N^+ \times N^+$),则三元组 $<a,l,v>$ 表示数值属性 a 的属性值是$[l,v]$.

定义 9 若满足 $\forall v \in V_c = [\alpha_1,\alpha_2,\cdots,\alpha_m]$, $m \in N^+$,则称 a_c 为类别型多值属性.如果 a_c 为类别型属性值,$a_c = <a,\alpha>$(其中 $<a,\alpha> \in I \times N^+$),则二元组 $<a,\alpha>$ 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 $<a,\upsilon,\mu>$、$<a,l,u>$ 和 $<a,\alpha>$ 称为项(Item),则 I 称为项集(ItemSet).

定义 10 若对于任意 $<a,\upsilon,\mu>$、$<a,l,u>$ 和 $<a,\alpha> \in \lhd i \rhd$,存在 $<a,q> \in t_j$,使得 $\upsilon \leqslant q \leqslant \mu$ 或 $q = [l,v]$或 $q = \alpha$ 成立,则称事务 t_j 支持 i.

定义 11 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 $\lhd i_l \rhd \bigcap \lhd i_r \rhd = \varnothing$.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 $c\%$.

3 算法描述

3.1 多值属性关联规则挖掘算法

本文引入了关键属性因子 KAF 和概念层因子 CHF 进行有选择性地频繁项集挖掘.在初始情况下,通过设置 KAF 和 CHF 值的大小,查询由关键因子构成的数据集,在以后的 k 频繁项集挖掘中,利用 CHF 因子对其进行初始化,将不同的频繁项集划分到不同的抽象层,便于进行可视化展示和各种应用分析.

算法 1 genFreqItemset()
输入:数据集 DB,最小支持度 minSup,KAF 因子和 CHF 因子.
输出:频繁项集 L.
genFreqItemset(DB,minSup,KAF,CHF)
(1)TID = get_key_value(DB,KAF,CHF);//TID 是根据 KAF 和 CHF 因子选择的关键数据集
(2)L_1 = get_frenqent_1_itemset(TID);
(3)freqInitial(频繁 1-项集,CHF);//初始化 1-itemset
(4)FOR($k = 2; L_{k-1} \neq \Phi; k++$)DO
(5) C_k = genCandidate(L_{k-1},CHF);//根据 CHF 因子值对 k-item 进行概念分层
(6) FOREACH 每个事务 $t \in$ TID DO
(7) FOREACH 每个候选项 $c \in C_k$ DO
(8) IF $c \in$ 事务 t THEN
$c.Count++$;//支持度计数增值
(9) $L_k = \{c \in C_k \mid c.\text{count} >= \text{minSup}\}$;
(10)return Sort($L = \bigcup L_k$);//将频繁项集按包含项的个数进行排序

genFreqItemset()算法的优点是引入了确定关键数据集和频繁项集概念分层的生成方法,利用 KAF$\{N_i, S_i, C_i\}$和 CHF$\{N_j, S_j, C_j\}$因子,设置数值型属性值集合 N_i,区间型属性值集合 S_i 和类别型属性的集合 C_i,并将不同的频繁项集划分到相应的概念层$\{N_j, S_j, C_j\}$上;在算法的初始执行阶段,依据 KAF 和 CHF 因子从数据集中选择的关键数据集,对频繁项集的项目进行约束,减少频繁项集的计算时间,同时也减少生成规则的计算时间;由于引进关键属性因子和概念层因子来定义查询的数据集,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.

算法 2 genRule()
输入:频繁项集 L,最小置信度 minConf,CHF 因子.
输出:关联规则 Rule.
genRule(L,minConf,CHF)
(1)FOR 每一个频繁 k-项集$f_k, k >= 2$ DO
(2) C_t = getSubset(f_k); //取每个频繁的子集
(3) IF C_t.Count > 0 THEN
(4) FOREACH 每个频繁子集 $c \in C_t$ DO
(5) Conf = TID.FindSupport(f_k)/TID.FindSupport(c);
(6) IF Conf >= minConf THEN
Rules(频繁子集→C_t.Remove(子集 c),CHF); //初始化规则前件和后件
(7) Ruleset = $\bigcup$Rules
(8)return RuleSet; //返回规则集合

genRule()算法利用关联规则概念分层的生成方法,通过设置 CHF 因子将规则的前件和后件进行概念层初始化,并将不同的规则前件和后件划分到相应的概念层$\{N_j, S_j, C_j\}$上,便于用户分析与研究不同概念层次的规则信息.

3.2 频繁项集可视化算法

算法 3 VFreqItems() //频繁项集可视化算法
输入:频繁项集 L,最小支持度 minSup,KAF 因子,CHF 因子.
输出:频繁项集可视化形式.
VFreqItems(L,minSup,KAF,CHF)
(1)freqItems = get_key_freqiItem(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 >= 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; //$k = item_{i.\text{count}}$
(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 $N_{(k+1)} \in L_{k+1}$ DO
(7) IF $N_{ki} \subset N_{(k+1)}$ THEN
(8) VF = VF$\bigcup\{N_{ki} \rightarrow N_{(k+1)}\}$; //$N_{ki}$指向其父节点 $N_{(k+1)}$

(9)return VF;

VFreqItems()算法的主要优点是:采用概念格的形式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与例化关系进行体现,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.

图 1 为算法的执行实例.第一步扫描频繁项集表,依据 KAF 和 CHF 因子大小选择频繁项集 F01 ~ F11;第二步将项集节点 F01 ~ 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 Fi 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}(后面的数字代表层数);第四步根据每个 Fi 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}→{F06,2},{F02,1}→{F08,2}.最后生成频繁项集可视化形式.

3.3　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,其中 RM 表示可视化的展示形式,CLN 表示规则前件与后件所允许跨层分析的层数.

算法 4　VRules()　//关联规则可视化算法

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

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 = '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)FOR　$R \in$ RuleSet && R.conf > = minConf DO　//R 表示规则

(7)　调用 Round(L_k, N_{ki})自动分配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

(8)　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}

(9)　　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 $N_{(k+CLN)j}$

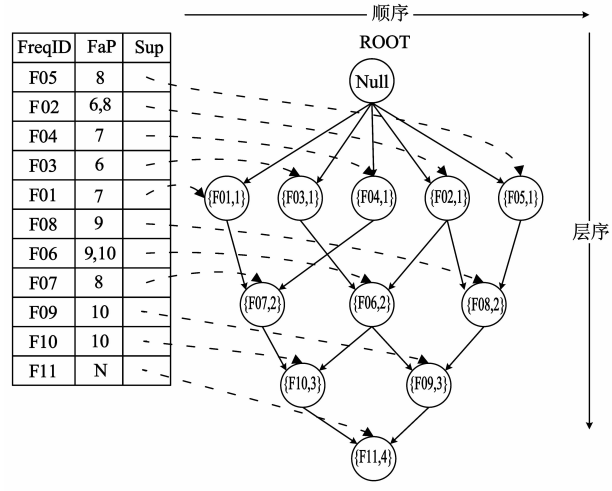
(10)　　IF $N_{ki} \in R$ && $N_{(k+CLN)_j} \in R$ THEN

(11)　　VR = VR∪{$N_{ki} \rightarrow N_{(k+CLN)j}$};

(12)return VR;

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.为了避免节点重叠,提高规则的友好性和展示效果,根据每层节点的特点和分布情况调用 Round(L_k, N_{ki})函数实现自动调整各个节点位置,使同层节点在同一个圆周上,使其具有较强的立体感.

图 2 为算法的执行实例.第一步扫描关联规则表,按 RM = "默认"和 CLN = 1 选择关联规则可视化形式;第二步将关联规则前件节点 L01 ~ L12 划分到不同层的圆周上,如图 3 - 1 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前件节点 Li 的 RHS 值(Right Hand Side,规则后件),将每层中规则前件节点指向规则后件节点.如前件节点 L04{5,7,8},生成规则 R:{04,2}→{05,2},R:{04,2}→{07,2},R:{04,2}→{08,3}(后面的数字代表项数).最后生成关联规则可视化形式.



FreqID	FaP	Sup
F05	8	-
F02	6,8	-
F04	7	-
F03	6	-
F01	7	-
F08	9	-
F06	9,10	-
F07	8	-
F09	10	-
F10	10	-
F11	N	-

图1　频繁项集可视化

4　多值属性关联规则可视化挖掘应用实例

4.1　多值属性关联规则挖掘过程

针对现有文献中的关联规则挖掘缺少统一的参数调整方法,导致用户无法通过设置其它参数来发现感兴趣的信息(例如关键字段的选择)等问题,文章结合概念格的多值背景理论建立了以支持度、置信度、关键属性因子和概念层因子为基础的参数调整策略,在整个挖掘过程中通过调整 Sup、Conf、KAF 和 CHF 参数的大小来挖掘相应的频繁项集和关联规则,如图 3 所示.通过参数模块 Pas 设置 KAF 因子和 CHF 因子值从输入的

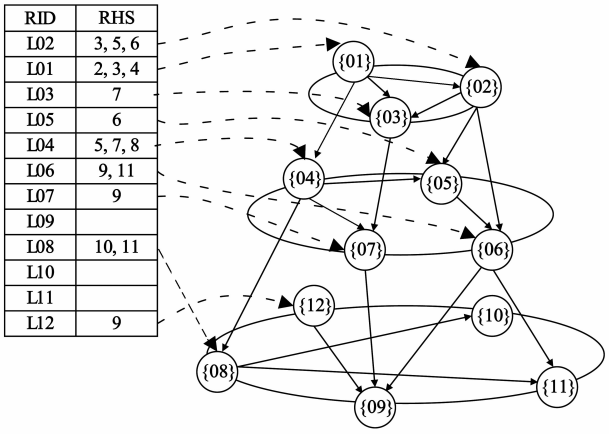


图2 关联规则可视化

数据集 DB 中选择关键属性值和不同的概念层,用户能够交互地调整 KAF 和 CHF 动态分析数据,最后设置 Sup 和 Conf 值挖掘频繁项集和关联规则.具体介绍如下:

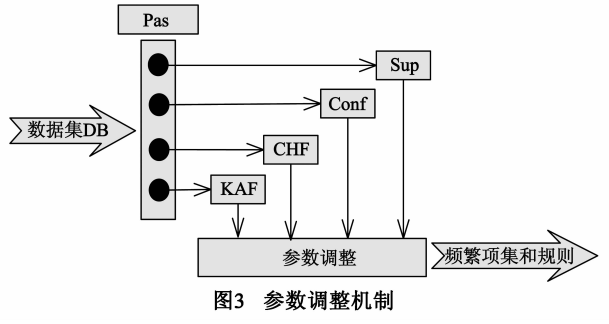


图3 参数调整机制

(1)多值属性值离散化,根据实际情况在关联规则挖掘之前对不同形式背景中多值属性字段进行离散化处理,依据定义 4、定义 5 和定义 6 对记录中的多值属性字段进行转换.本文选择全员人口数据库中的四个属性值(年龄,户口性质,世代间隔,管理地)如表 1 所示,对数值型字段“年龄”和“世代间隔”,类别型字段“户口性质”和“管理地”进行离散化操作,得到结果如表 2、表 3、表 4 和表 5 所示.离散化完成后,结果如表 6 所示.

(2)设置 KAF 因子,主要是选择数据库记录中关键字段.挖掘有趣规则最关键的一点是如何为算法选择适当的属性值,因此在关联规则挖掘过程中,用户通过设置 KAF$\{N_{ki}, S_{ki}, C_{ki}\}$因子来选择记录(项目)集合 $A=\{a_1, a_2, \cdots, a_k, \cdots, a_n\}$(其中,$k$ 表示关键属性,i 表示集合项数)中的关键属性项进行分析,其中 N_{ki} 表示数值型属性值集合,S_{ki} 表示区间型属性值集合,C_{ki} 表示类别型属性的集合,其中 $i\in N^+$.如:KAF{(年龄),(间隔),(户口性质)}.

(3)设定 CHF 因子,主要是将数据库中的字段项按 CHF$\{N_{hj}, S_{hj}, C_{hj}\}$因子值(其中,$h$ 表示概念层,用于区分关键属性因子,j 表示属性所属层级)进行概念分层初始化,其中 N_{hj} 表示数值型属性值的概念层数,S_{hj} 表示区间型属性值的概念层数,C_{hj} 表示类别型属性的概念层数,其中 $j\in N^+$,如 CHF{3,3,3}.如图 4 所示,为由$\{\{t_1$,文化$\},\{t_2$,初级$\},\{t_3$,小学$\},\cdots\}$等构成的概念层形式,其中$\{\{t_2$,初级$\},\{t_4$,中级$\},\{t_5$,高级$\}\}$具有相同(第二层)的层关系;$\{\{t_6$,小学$\},\cdots,\{t_9$,中专$\},\cdots,\{t_{12}$,本科$\}\}$具有相同(第三层)的层关系.

表 1 人口数据表

人口编号	年龄	户口	世代间隔	
1307001	22	其它	[15 – 19]	农村
1307002	28	非农业	[25 – 29]	城镇
1307003	24	农业	[20 – 24]	农村

表 2 年龄属性

年龄	整数值
[20 – 24]	1
[25 – 29]	2
[30 – 34]	3

表 3 户口性质属性

户口	整数值
农业	1
非农业	2
其它	3

表 4 世代间隔属性

间隔	整数值
[15 – 19]	1
[20 – 24]	2
[25 – 29]	3

表 5 管理地属性

管理地	整数值
农村	0
城镇	1

表 6 人口数据表(离散后)

人口编号	年龄	户口	世代间隔	
1307001	1	3	1	0
1307002	2	2	3	1
1307003	1	1	2	0

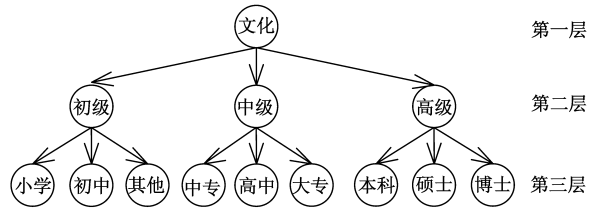


图4 概念分层

(4)完成前三步操作后,设定最小支持度 Sup、最小置信度 Conf 的大小进行多值关联规则挖掘.

4.2 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.

4.2.1 育龄妇女世代间隔频繁项集可视化

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,本节主要通过调用 VFreqItem()算法来发掘它们之间的频繁模式,设置 KAF{{年龄},{间隔},{文化,地区,户口性质}}、CHF{1,1,1}因子和 minSup > = "45.00"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.本文以概念格的形式将频繁项集中的数据项组织起来,并利用渐变颜色表示不同支持度的频繁项集,其中圆形节点表示每个频繁项集,连接线表示项集之间的关系;频繁项集的支持度越大,该项集的颜色越深,如图 5 中图例所示.(注:本图原图为彩色,为便于印刷,我们将其转换为了黑白灰度图.灰度越高表示频繁项集的支持度越大)

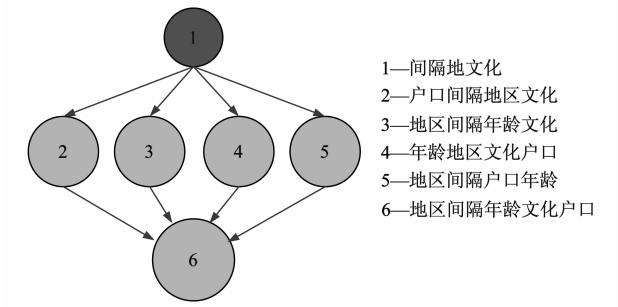


图5　频繁项集可视化

4.2.2 育龄妇女世代间隔关联规则可视化

本节对 4.2.1 节所挖掘的频繁项集进行多模式关联规则可视化挖掘,调用关联规则可视化算法 VRules()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关联规则可视化展示,如图 6 为了黑白所示(注:本图原图为彩色,为便于印刷,我们将其转换灰度图.灰度越高表示置信度越大.).同 4.2.1 节,我们用规则前后件连接线的线型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.

多对多模式是多值属性关联规则可视化挖掘的重要展示方式,如设置 KAF{{年龄},{间隔},{文化,地区,户口性质}}、CHF{1,1,1}因子和 minConf > = "35.50"来分析多属性之间的相关关系,其中图 7 和图 8 分析展示多对多、多对一模式的关联规则,另外从图 7 中我们能够得到如表 7 所示的人口规则信息.

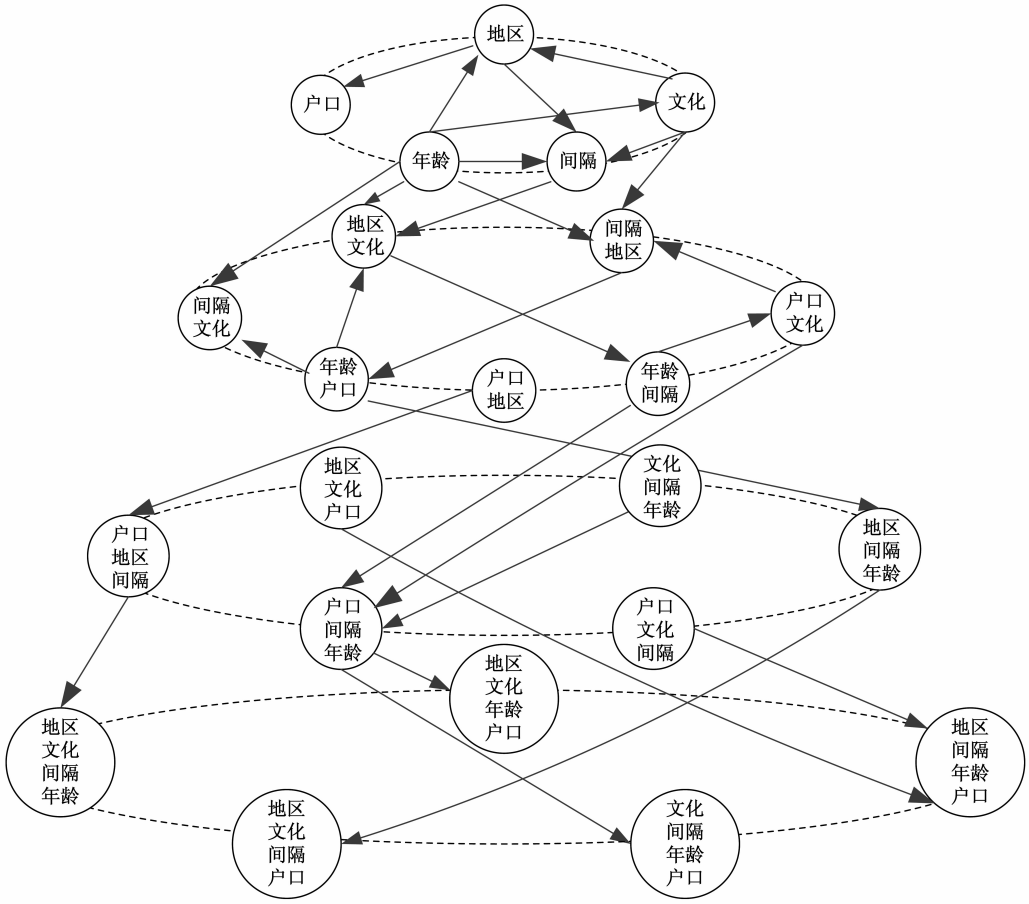


图6　关联规则可视化展示

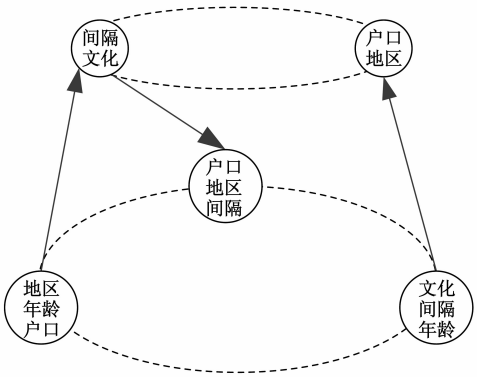


图7 多对多模式

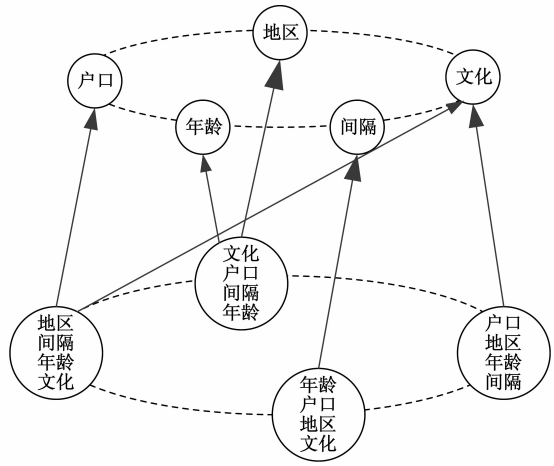


图8 多对一模式

表 7 多对多模式

规则前件	规则后件	置信度
间隔,文化	地区,年龄,户口	87.25%
文化,间隔,年龄	户口,地区	65.38 %
户口,地区,间隔	间隔,文化	77.56 %

概念分层的多值属性关联规则展示形式也是关联规则可视化挖掘过程不可或缺的组成部分.本文通过对多值属性数据进行概念分层处理,帮助用户进行基于概念分层的关联规则可视化挖掘,从而发现更具有价值的信息.通过设置 KAF{{地区},{文化},{年龄,间隔}}、CHF{3,3,3}(或 CHF{2,2,2})和置信度 > = "25.00"来以概念分层的形式展示年龄、间隔、地区和间隔之间不同抽象层之间的关系,如图 9 所示.

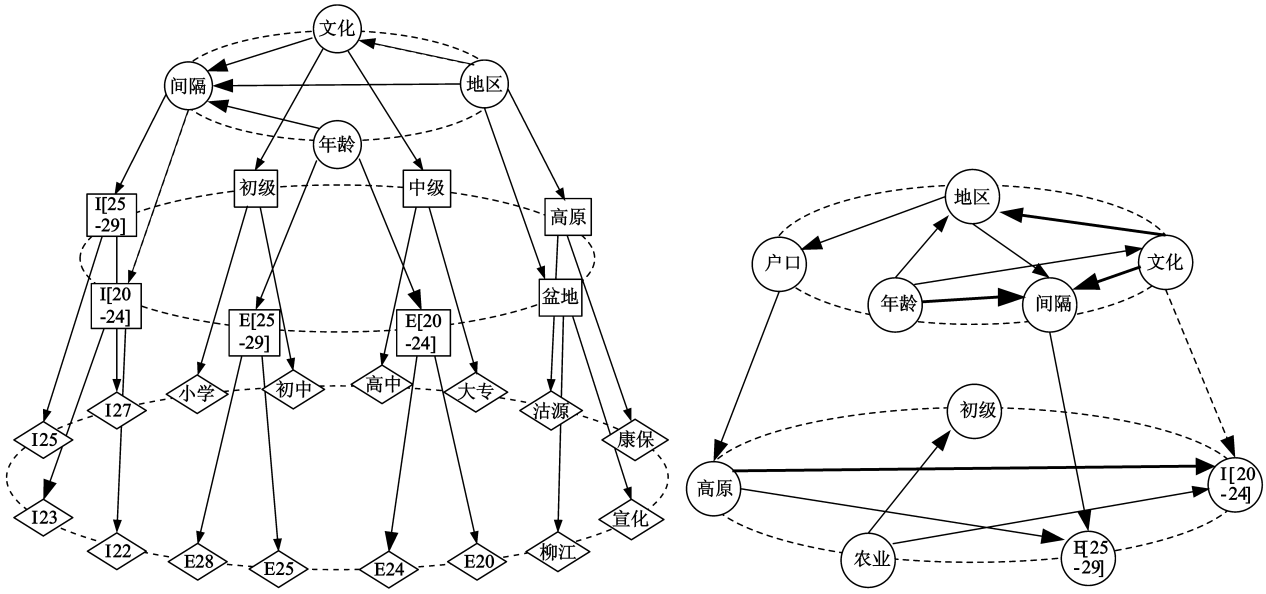


图9 概念分层展示模式

5 实验分析

5.1 多值属性关联规则可视化效果比较

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 图以及平行坐标系的规则可视化[6,10],相比具有以下优点:

方便用户动态分析不同类型字段项之间的频繁模式和关联关系;规则前后件信息区分明显,有效避免出现规则重叠的现象;展示形式具有表现力较强、结构严谨和非模糊的特点;同时实现了频繁项集与多模式关联规则可视化展示.

5.2 多值属性关联规则算法性能比较

为了验证本文改进算法的正确性、高效性和其它各种性能,本文采用以下实验环境:测试程序在 Windows Server 2008 系统上运行,主频 3.10GHz,4GB RAM,VS2008,数据库系统为 ORACLE 10G,算法实现均采用

C# 语言,测试数据为某省全员人口数据.经过离散化处理后,所生成数据集的具体控制参数含义及其缺省值详见表 8.

表 8　测试数据的相关参数

参数符号	具体含义	设置大小
|TID|	记录个数	100 ~ 324000
I	项的数目	5 ~ 15
A	项的属性个数	1 ~ 5
F	频繁项集个数	300
V	最大频繁项集平均长度	2 ~ 10
H	概念分层数	3 ~ 4

本文主要从记录数量变化、支持度大小变化和项数目变化 3 个方面,对 Apriori 算法和改进算法进行了具体的实验比较和分析,具体结果如图 10、图 11 和图 12 所示.

(1)记录数量变化.当同时取最小支持度 5.0,项目数量为 10 个时,记录数从 100 条逐步增至 300*K* 条.由图 10 可以看出,在同等条件下,改进算法的执行速度明显快于 Apriori 算法.在记录数量相对较小的情况下,二者区分效果不够明显,但是随着数据量的不断增加,改进算法的执行效果逐步显示出明显的优势,当记录数为 300K 时,改进算法执行时间为 1.684s,只有 Apriori 算法执行时间 5.606s 的 30%,性能提高了 3.3 倍.

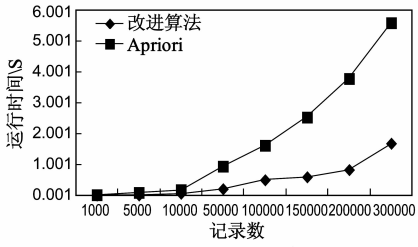


图10　记录数量变化

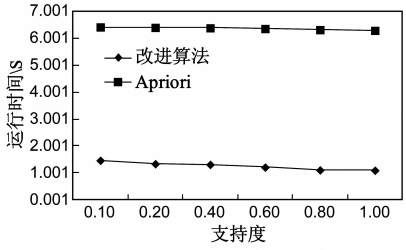


图11　支持度大小变化

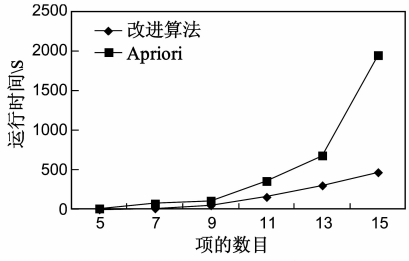


图12　项的数目变化

(2)支持度大小变化.当取记录数为 324K 条,项目数量为 10 个时,最小支持度从 0.1 逐步增至 1.0.由图 11 可以看出,在数据记录数相同的情况下,随着最小支持度的不断增加,两种算法的运行时间随之减小,但在整个变化过程中改进算法的平均执行时间为 1.25s,而 Apriori 的平均执行时间为 6.3s,性能约提高了 5 倍.

(3)项数目变化.当取记录数为 324K 条,最小支持度为 5.0 时,项的个数从 5 个逐步增至 15.由图 12 的实验结果可知,在项的数目相对较少的情况下,二者区分效果不够明显,但是伴随数据项数量的不断增加,改进算法的执行效率显然比 Apriori 算法具有明显优势,当数据项增至 15 个时,Apriori 算法的执行时间是 1945.987s,而改进算法执行时间是 469.987s,性能约提高了 4 倍多.

以上实验结果表明,多值属性关联规则挖掘中,在相同条件下,引入 KAF 和 CHF 因子的 Apriori 算法的执行速度比传统的 Apriori 算法快,显著地提高了挖掘性能.

6　结束语

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挖掘方法,该算法可以有效地展示一对一、一对多、多对一、多对多和概念分层的关联规则可视化展示形式;运用概念格理论对其进行了具体分类并建立较为完整的挖掘过程参数调策略,采用基于 KAF 因子和 CHF 因子的 Apriori 改进算法进行多值属性关联规则挖掘,执行速度与 Apriori 算法相比有较大提高,具有更好的挖掘效率和性能.最后,以某省全员人口数据对算法进行了具体实现和分析,实验结果表明所提出的关联规则可视化表现形式具有良好的可视化展示效果,改进算法具有更好的挖掘性能.在下一步的研究中,我们将针对如何利用频繁项和关联规则之间的语义联系、应用背景,将频繁项集和规则转换为领域知识等问题进行改进.

参考文献

[1] Wang D X, Xie Q. Analysis of association rule mining on quantitative concept lattice [A]. Artificial Intelligence and Computational Intelligence, LNCS7530 [C]. Berlin: Springer-Verlag, 2012. 142 – 149.

[2] BayVo, Bac Le. Interestingness measures for association rules: combination between lattice and hash tables [J]. Expert Systems with Applications, 2011, 38(9): 11630 – 11640.

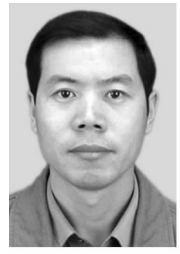
[3] Li D C, Zhang M. A new approach of self-adaptive discretization to enhance the Apriori quantitative association rule mining [A]. Proceedings of the 2012 Second International Conference on Intelligent System Design and Engineering Application[C]. Washington, DC: IEEE Computer Society, 2012. 44 – 47.

[4] 刘波,潘久辉.基于频繁模式图的多维关联规则挖掘算法研究[J].电子学报,2007,35(8):1612 – 1616.
Liu Bo, Pan Jiu-hui. Research of algorithms based on a frequent

pattern graph for mining multidimensional association rules [J]. Acta Electronica Sinica, 2007, 35(8): 1612 – 1616. (in Chinese)
[5] Bal M, Bal Y, Ustundag A. Knowledge representation and discovery using formal concept analysis: an HRM application[A]. Proceedings of the World Congress on Engineering[C]. London: Newswood, 2011. 1068 – 1073.
[6] Cassio M, Legrand B. Extracting and visualising tree-like structures from concept lattices[A]. Proceedings of the 2011 15th International Conference on Information Visualisation [C]. Washington, DC: IEEE Computer Society, 2011. 261 – 266.
[7] 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.
[8] Michael H, Chelluboina S. Visualizing association rules in hierarchical groups[A]. Interface 2011: Statistical, Machine Learning, and Visualization Algorithms [C]. North Carolina: SAS Institute, 2011. 1 – 11.
[9] Dario B, Cristine D. Visual mining of association rules[A]. Visual Data Mining: Theory, Techniques and Tools for Visual Analytics, LNAI 6208 [C]. Berlin: Springer-Verlag, 2008. 103 – 122.
[10] Alatas B, Akin E, Karci A. MODENAR: multi-objective differential evolution algorithm for mining numeric association rules [J]. Applied Soft Computing, 2008, 8(1): 646 – 656.
[11] Pachón Álvarez V, Mata Vázquez J. An evolutionary algorithm to discover quantitative association rules from huge databases without the need for an a priori discretization [J]. Export Systems with Applications, 2012, 39(1): 585 – 593.
[12] Martínez-Ballesteros M, Riquelme J. Analysis of measures of quantitative association rules[A]. Proceedings of the 6th international conference on Hybrid artificial intelligent systems [C]. Berlin: Springer Verlag, 2011. 6679. 319 – 326.
[13] 耿生玲,李永明,刘震.关联规则挖掘的软集包含度方法[J].电子学报,2013,41(4):804 – 809.
GENG Sheng-ling, LI Yong-ming, LIU Zhen. An Approach to Association Rules Mining Using Inclusion Degree of Soft Sets [J]. Acta Electronica Sinica, 2013, 41(4): 804 – 809. (in Chinese)
[14] Martinez-Ballesteros M, Riquelme J. Analysis of measures of quantitative association rules[A]. Proceedings of the 6th International Conference on Hybrid Artificial Intelligent Systems [C]. Berlin: Springer-Verlag, 2011. 319 – 326.
[15] Ganter B, Wille R. Formal Concept Analysis: Mathematical Foundations [M]. Berlin: Springer Verlag, 1999. 17 – 35.
[16] Nguyen TT, Hui SC, Chang K. A lattice-based approach for mathematical search using formal concept analysis [J]. Expert Systems with Applications, 2012, 39(5): 5820 – 5828.

作者简介

郭晓波 男,1986 年生于河北栾城.硕士研究生,研究方向为数据挖掘与智能信息处理.



赵书良(通信作者) 男,1967 年生于河北献县,教授,博士,博士研究生导师,主要研究方向为智能信息处理.
E-mail: zhaoshuliang@sina.com