

## 基于 S-C 元图的关联规则表示方法

陈 敏<sup>1,2,3</sup> 赵书良<sup>1,2,3</sup> 郭晓波<sup>1</sup> 柳萌萌<sup>1,2,3</sup> 李晓超<sup>1,2,3</sup>

(河北师范大学数学与信息科学学院 石家庄 050024)<sup>1</sup>

(河北省计算数学与应用重点实验室 石家庄 050024)<sup>2</sup> (河北师范大学移动物联网研究院 石家庄 050024)<sup>3</sup>

**摘 要** 针对现有的关联规则可视化形式中,面向专家用户而忽略了普通用户的感知能力,以及当规则数量增多时,边和节点的代表元素易发生重叠,缺乏有效的展现形式,降低了可读性等问题,提出了一种新的基于 S-C 元图形式的一对一、一对多、多对一、多对多的关联规则可视化表示方法。首先给出了 S-C 元图的基本定义和以 S-C 元图展示关联规则的模型;接着给出了 S-C 元图的性质和推导过程;最后以某省全员人口数据为基础,结合前注意加工理论与格式塔理论,以 S-C 元图和纺锤体相结合的形式对多模式的关联规则进行可视化展示,并分析了展示效果。实验结果表明,所提出的可视化方法具有良好的展示效果。

**关键词** 元图,关联规则,可视化,格式塔理论,前注意加工理论

**中图法分类号** TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2014.08.056

### Visualization of Association Rules Based on S-C MetaGraph

CHEN Min<sup>1,2,3</sup> ZHAO Shu-liang<sup>1,2,3</sup> GUO Xiao-bo<sup>1</sup> LIU Meng-meng<sup>1,2,3</sup> LI Xiao-chao<sup>1,2,3</sup>

(Mathematics and Information Science College, Hebei Normal University, Shijiazhuang 050024, China)<sup>1</sup>

(Hebei Key Laboratory of Computational Mathematics and Applications, Shijiazhuang 050024, China)<sup>2</sup>

(Institute of Mobile Internet of Things, Hebei Normal University, Shijiazhuang 050024, China)<sup>3</sup>

**Abstract** Considering the problems aroused by the traditional association rules visualization approaches which are orienting to expert users while ignoring the normal user perception, even more when the number of rules increases, particularly prone to overlap among edges and nodes, as well as result in reducing the performance and readability of rule representation, this paper proposed a novel form of visualization display method based on S-C metagraph to show one-to-one, one-to-many, many-to-many association rules. Firstly, gave the basic definition of S-C metagraph and the model showing association rules using S-C metagraph. Then gave the properties and derivation process of S-C metagraph for visualizing association rules. Finally, with the help of experimental data obtained from demographic data of a province, combining with the Preattentive Processing Theory and Gestalt Theory, we illustrated multi-mode association rules in the combination form of S-C metagraph and spindle, and analyzed the effect of the visualization display. The realistic application analysis and experimental results turn out that the proposed visualization method has excellent visual effects.

**Keywords** MetaGraph, Association rules, Visualization, Gestalt theory, Preattentive processing theory

## 1 引言

关联规则可视化表示是一种流行的数据挖掘技术,其目的是从海量数据中展现出属性间潜在的、有价值的频繁模式和相关关系。通常用户需要分析大量的关联规则来获取有价值的信息和内容,这对关联规则的可视化与交互式分析工具提出了迫切需求。近年来涌现出诸多关联规则可视化形式和交互式探索系统与工具;Buono 等<sup>[1]</sup>提出一个可视化数据挖掘框架,利用基于图的方法宏观展示关联规则挖掘结果,并采用平行坐标技术微观探究特殊的规则集合,然而该方法却无

法避免规则数量庞大时节点或边的重合。Ertek 等<sup>[2]</sup>设计了一种基于图的可视化形式,用图形的大小和颜色代表支持度和置信度,但仅适用于展示单维单层布尔型关联规则,无法实现多模式展现形式。Hahsler 等<sup>[3]</sup>改进了基于矩阵的可视化方法,利用嵌套层次分组,使用户逐层探索关联规则,在一定程度上适应了大型规则集,但是无法向用户展示所有关联规则的总体特征。Guimei Liu 等<sup>[4]</sup>设计了一个交互式关联规则可视化系统 AssocExplorer,用户可观察所有关联规则的总体特征,通过缩放和过滤操作选择需要的关联规则以及查看所选关联规则的详细信息,然而该系统操作与展示效果较为专

到稿日期:2013-10-09 返修日期:2013-12-13 本文受国家自然科学基金项目(71271067),国家社会科学基金项目(13BTY011),河北师范大学数学与信息科学学院硕士基金(201301)资助。

陈 敏(1988—),女,硕士生,CCF 学生会会员,主要研究方向为数据挖掘、智能信息处理,E-mail:cm316029906@126.com;赵书良(1967—),男,博士,教授,博士生导师,主要研究方向为智能信息处理,E-mail:zhaoshuliang@sina.com(通信作者);郭晓波(1986—),男,硕士生,主要研究方向为数据挖掘;柳萌萌(1988—),女,硕士生,主要研究方向为数据挖掘、智能信息处理;李晓超(1987—),男,硕士生,主要研究方向为数据挖掘、智能信息处理。

业,不能够很好地面向普通用户。Bornelöv等<sup>[5]</sup>将包含项的规则映射到一个圆环图形上,通过一条由蓝逐渐变黄的线段将两个规则连接在一起,规则的前件用线段的箭头指向后件。该方法可以直观地显示不同项之间的关系,并且能够突出重点信息,但不能更好地表示规则的支持度和置信度。Cinicioglu等<sup>[6]</sup>利用图论的思想,用点代表一个项(或规则的前、后件),用边代表规则的关系,其中用边的颜色表示不同的支持度,用边的宽度代表规则的置信度,能够非常直观地表示各种关联规则,但是不适合可视化包含大量属性的规则。郭晓波等<sup>[7]</sup>提出一种新的基于概念图的关联规则知识表示方法,该方法将关联规则所涉及的概念属性信息及相互关系转换成概念结点和关系节点,并能够准确、简洁且充分地展示规则信息,但是缺乏对元知识、概念图的信息聚类模式的展示。Bruzzeze等<sup>[8,9]</sup>对各种关联规则可视化形式进行了综合比较与分析,该方法一般仅能展示较少的关联规则,交互性和可用性较差,无法满足普通用户的需求。

随着面向普通用户的信息可视化技术的发展<sup>[10,11]</sup>,用户的感知能力在信息可视化中扮演的角色越发重要。心理学领域著名的前注意加工理论(Preattentive Processing Theory)和格式塔理论(Gestalt Theory)为实现信息可视化的价值提供了有利的理论支撑:前注意加工理论解释了一些视觉特征(如:颜色、线条方向、线条长度和宽度、闭环等)可以被低层视觉系统快速、精确地感知;格式塔理论解释了高层视觉系统试图理解图像时所遵循的原则,即距离较近的、相似的、连续的、对称的或者能形成闭环的元素更倾向于被识别为一个整体,相对较小的元素易被识别为较大元素的背景<sup>[12]</sup>。Peterson等<sup>[13]</sup>通过实验证明了利用格式塔理论的相似性原则分组可以提高视觉工作记忆(Visual Working Memory)的性能。Rusu等<sup>[14]</sup>利用格式塔的闭合原则,在图可视化中边交叉的地方引入“缺口”,增强了图的可读性。Healey等<sup>[15]</sup>总结了视觉前注意加工理论,介绍了视觉注意在感知突显、预测注意力和吸引注意力中的应用。Michalski等<sup>[16]</sup>通过实验证明了视觉前注意加工机制扮演了重要角色,提倡将其应用于人机交互领域;同时证明了颜色前注意视觉加工在空间结构相当复杂的情况下作用不显著。已有一些关联规则可视化系统<sup>[2-4]</sup>利用了元素颜色和线条的特征以及邻近性原则,但是不适用于普通用户理解和使用,且无法避免大量关联规则展示时所造成的视觉负担。

已有的关联规则可视化方法大都存在以下不足:无法实现多模式关联规则可视化展示;关联规则数量多时会造成节点或边的重叠;用户友好性较差,不适用于缺乏领域知识的普通用户使用。

针对上述问题,本文提出一种新的基于元图形式——S-C元图(S代表Support和C代表Confidence)的关联规则多模式表示方法。该方法能够对规则的属性集合进行有效合理的组织,将规则前件作为入度集合、规则后件作为出度集合,利用元素的颜色和线条特征以及格式塔理论和前注意加工理论对一对一、一对多、多对一、多对多的关联规则进行可视化展示与分析。所提算法能够有效地识别海量数据中频繁出现的重要(关键)属性,同时避免了线条或者节点的重合,展示效果具有较好的可读性。最重要的是,它极大地降低了普通用户的感知负担,便于用户从不同角度分析和利用挖掘结果。同时,基于某省全员人口数据,以S-C元图结合纺锤体的形式对

算法进行了具体实现和分析,以证明本文方法的可行性。

## 2 关联规则的S-C元图表示方法

### 2.1 S-C元图

元图是图论领域的新兴研究课题,其利用人类很强的视觉处理能力,用信息进行可视化展示,能够提高可视化过程的效率<sup>[17]</sup>。因此,元图继承了图的优势,不仅在图形可视化功能上优于其他图,而且具有很强的形式化分析能力。

Blanning等<sup>[18]</sup>分析了简单图、有向图、超图、分层图和Petri网的局限性,提出一种新的图结构——元图。元图发展了有向图和超图的优势,可展示集合和集合之间的关系,这也为实现多模式关联规则可视化提供了有利条件和理论支撑。目前,元图已经被广泛应用于模型管理、数据和规则管理等领域中。谭政华等<sup>[19]</sup>提出了模糊元图的概念及其代数结构,并分析了其在规则管理中的应用。Dashore等<sup>[20]</sup>将模糊元图应用于基于规则的股票市场决策系统,取得了良好的效果。为了更好地对关联规则进行可视化表示,本文给出基于S-C元图的关联规则表示方法。

**定义1(元图)** 元图 $S$ 是定义在生成集 $X=\{x_i, i=1, 2, \dots, N\}$ 上的一个序对,即 $S=\langle X, E \rangle$ 。其中 $E=\{e_j, j=1, 2, \dots, N\}$ 表示元图 $S$ 中所有边组成的集合,对于任意的 $e_j \in E$ 是由生成集 $X$ 的两子集 $V_j, W_j \subseteq X$ (其中 $V_j \cup W_j \neq \emptyset$ )组成的序对,记 $e_j=\langle V_j, W_j \rangle$ ,分别称 $V_j, W_j$ 为 $e_j$ 的入点集(Invertex)和出点集(Outvertex)。

为了更好地实现基于S-C元图的关联规则可视化任务,本文对元图进行了改进,提出S-C元图的定义:

**定义2(S-C元图)** S-C元图 $S$ 是定义在生成集 $X=\{x_i, i=1, 2, \dots, I\}$ 上的一个三元组,即 $S=\langle X, E, W \rangle$ 。其中 $E=\{e_k, k=1, 2, \dots, K\}$ 为边集,记 $e_k=\langle V_k, W_k \rangle$ 由 $X$ 的子集作为端点集,满足 $V_k, W_k \subseteq X$ 且 $V_k \cup W_k \neq \emptyset, V_k \cap W_k = \emptyset$ ,称 $V_k, W_k$ 分别为 $e_k$ 的入点集和出点集。 $W=\{i_1, i_2, i_3, \dots, i_n\}, i_r=\{Supp_r, Conf_r\}, 1 \leq r \leq n, n$ 是关联规则的数目, $Supp_r$ 表示第 $r$ 个关联规则的支持度, $Conf_r$ 表示第 $r$ 个关联规则的置信度。

### 2.2 关联规则的S-C元图表示形式

S-C元图不仅能够展示包含单值属性的关联规则,而且具有良好的多值属性关联规则表示效果,并且便于获得对属性值重要性的评估。根据关联规则前件和后件的属性维度,可将关联规则分为一对一(one to one,  $x y$ )、一对多(one to many,  $x y, \dots, z$ )、多对一(many to one,  $x, \dots, y z$ )与多对多(many to many,  $x, \dots, y z, \dots, w$ )4种模式。

S-C元图是表示规则前件集合、规则后件集合以及相关信息的三元关系,为实现多模式的关联规则表示提供了有利条件。下面给出S-C元图关联规则表示的相关定义和数学模型。

**定义3(元值库)** 定义元值库 $K$ 为关联规则挖掘的数据预处理过程中所选择的属性值集合。即 $K=\{k_i, i=1, 2, \dots, N\}$ 为关联规则的S-C元图表示形式 $S=\langle X, E, W \rangle$ 中的生成集 $X$ 。若在关联规则挖掘的数据预处理过程中共选择了 $Y$ 个属性,且每个属性有 $y_i$ 个值,则有 $I=y_i * Y$ 成立。

**定义4(元值)** 元值库 $K$ 中的任一元素 $k_i$ 为元值,代表关联规则的前件或后件中的某一属性值,是S-C元图的最小单位。

**定义5(元属性)** 已知元值库 $K$ 的子集 $l=\{k_1, \dots, k_l\}$

$\subseteq K$  且  $k_i \in l$  属于同一属性, 则称该属性为元属性。给定元值库  $K = \{k_i, i=1, 2, \dots, N\}$ , 定义元值库  $K$  上的元属性  $T$ , 若  $k_1, k_2, \dots, k_i \in K$  且属于同一属性, 则记元属性为  $T = k_1 \Delta k_2 \Delta \dots \Delta k_i$ 。

**定义 6(元边)** 元边  $F$  是连接一条关联规则前件和后件的边的集合。假设某条关联规则的前件中包含  $M$  个元属性, 后件中包含  $N$  个元属性, 连接前件属性的  $M$  条边与连接后件属性的  $N$  条边交于一点  $P$ , 则该点保存此关联规则的支持度  $Supp$  和置信度  $Conf$ 。

**定义 7(元规则)** 关联规则的 S-C 元图表示形式中, 元值、元属性、元边表示的一条关联规则称为元规则, 根据模式的不同, 分为一对一规则、一对多规则、多对一规则、多对多规则。

给定元值库  $K = \{k_i, i=1, 2, \dots, I\}$ , 定义  $K$  的幂集  $P(K)$  到  $P(K)$  的映射  $\phi: P(K) \rightarrow P(K)$  表示关联规则前件与后件的关系。若对于任意的  $\{k_a, k_b, \dots, k_f\} \in P(K), \{k_j, k_k, \dots, k_t\} \in P(K)$ , 满足  $\{k_a \in T_a, k_b \in T_b, \dots, k_f \in T_f\} \in P(K), \{k_j \in T_j, k_k \in T_k, \dots, k_t \in T_t\} \in P(K)$ , 其中  $T_a \cap T_b \cap \dots \cap T_f \cap T_j \cap T_k \cap \dots \cap T_t = \emptyset$ , 即元属性不存在交集, 且  $\phi(\{k_a, k_b, \dots, k_f\}) = \{k_j, k_k, \dots, k_t\}$ , 则称  $\{k_a, k_b, \dots, k_f\}$  为规则前件,  $\{k_j, k_k, \dots, k_t\}$  为规则后件。

元规则展示规则信息中所有元值所属的元属性, 规则前件所属的元属性展示在第一层, 规则后件所属的元属性展示在第二层, 元边连接前件元值和后件元值, 并保存该规则的支持度和置信度。

基于 S-C 元图的一对一、一对多、多对一、多对多形式的关联规则表示分别如图 1—图 4 所示。

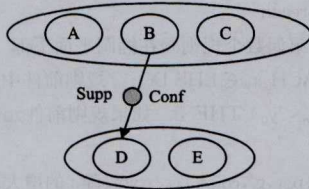


图 1 一对一关联规则的 S-C 元图表示

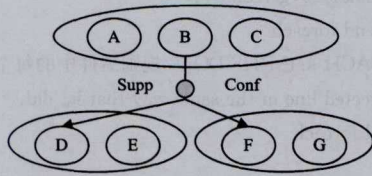


图 2 一对多关联规则的 S-C 元图表示

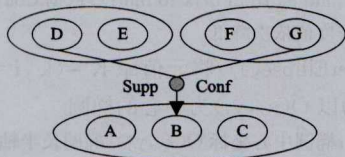


图 3 多对一关联规则的 S-C 元图表示

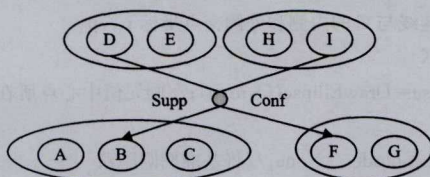


图 4 多对多关联规则的 S-C 元图表示

为避免展示多条关联规则时边的交叉现象, 本文以纺锤体为载体展示基于 S-C 元图的关联规则。图 5 展示了 4 种模式的关联规则, 黑线代表一对一关联规则:  $\{Y=d\} \Rightarrow \{U=l\}$  ( $Supp=10\%, Conf=30\%$ ); 灰线代表一对多关联规则:  $\{Z=f\} \Rightarrow \{X=a, V=o\}$  ( $Supp=30\%, Conf=15\%$ ); 虚线代表多对一关联规则:  $\{Z=h, W=j\} \Rightarrow \{X=a\}$  ( $Supp=10\%, Conf=15\%$ ); 点线代表多对多关联规则:  $\{X=a, Y=c\} \Rightarrow \{Z=h, V=m\}$  ( $Supp=20\%, Conf=30\%$ )。

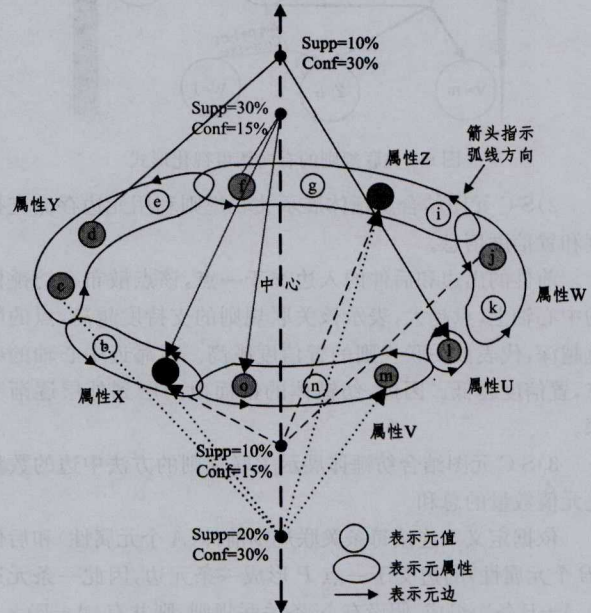


图 5 S-C 元图结合纺锤体的关联规则可视化形式

如图 5 所示, S-C 元图结合纺锤体的关联规则表示形式整体呈纺锤体结构, 元值均匀分布成椭圆型, 并可识别属于同一元属性的元值集合, 元值灰度值越大代表该元值更重要。每条元规则的元边各自形成感知上的锥面, 且元边颜色各不相同, “中心”上部呈暖色系, “中心”下部呈冷色系, 远离“中心”, 颜色逐渐加深, 代表置信度增大。垂直辅助轴上分布着大小不一的圆点, 圆点越大, 代表支持度越大。以关联规则  $\{X=a, Y=c\} \Rightarrow \{Z=h, V=m\}$  ( $Supp=20\%, Conf=30\%$ ) 为例, 元边为点线, 与虚线元边相比, 代表更大的置信度; 中心轴上的圆点较小, 表示支持度较小; 属性中  $a$  的灰度值比  $h$  大,  $h$  的灰度值比  $c, m$  大, 说明  $a$  比  $h$  更重要,  $h$  比  $c, m$  更重要。从图 5 可以看出,  $a$  与 3 条关联规则相关, 而  $h$  与 2 条关联规则相关,  $c$  和  $m$  仅与 1 条关联规则相关。下节将对图 5 进行详细分析。

### 3 S-C 元图结合纺锤体展示关联规则的特点分析

S-C 元图结合纺锤体展示关联规则克服了有向图可视化形式的缺点, 考虑了前注意加工理论和格式塔理论对感知系统的影响, 现将其特点总结如下:

1) S-C 元图结合纺锤体展示关联规则避免了二维平面上边的交叉。

如图 5 所示, 保存支持度和置信度信息的点散布在中心轴上, 因此单条关联规则的元边可依附于纺锤体的一个锥面, 而各条关联规则的元边与纺锤体中心平面的夹角不同, 所以各条元边所形成的锥面不会重叠, 各条关联规则的元边不会交叉。图 6 利用有向图表示与图 5 相同的 4 条关联规则, 边

的颜色代表置信度,颜色越深,置信度越高;边的粗细代表支持度,边越粗,支持度越高。一旦增加一条诸如 $\{Y=c, V=o\} \Rightarrow \{Z=h, W=j\}$  ( $Supp=20\%, Conf=15\%$ )的关联规则,就会出现交叉边,不易识别,给读者带来视觉负担。

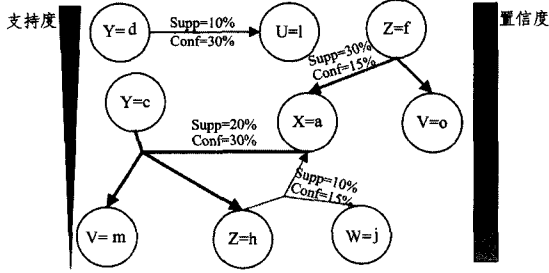


图6 关联规则的有向图可视化形式

2) S-C元图结合纺锤体展示关联规则利用元边存储支持度和置信度信息。

前件的出边和后件的入边交于一点,该点散布在纺锤体的中心轴上,点越大,表示该关联规则的支持度越高,点的颜色越深,代表该关联规则的置信度越高。越靠近中心轴的中点,置信度越低。因此,纺锤体的锥面由内层到外层逐渐变深。

3) S-C元图结合纺锤体展示关联规则的方法中边的数量是元值数量的总和。

依据定义5连接每条关联规则前件( $A$ 个元属性)和后件( $B$ 个元属性)的边交于一点 $P$ 形成一条元边,因此一条元边由 $A+B$ 条边组成,假设有 $N$ 条关联规则,则共有 $(A+B) * N$ 条边。因每条元边依附一个锥面,所以边的数目大,并不会造成平面交叉。依据格式塔理论,距离较近的、相似的、连续的、对称的或者能形成闭环的元素更倾向于被识别为一个整体。某条关联规则边的数目多且边的颜色相同,反而有利于读者的感知系统将此元边置于锥面背景下。

4) S-C元图结合纺锤体展示关联规则可描述属性的重要程度。

在关联规则中出现频率越高的元值颜色越深。图5中,元值 $a$ 出现在3个关联规则中,因此颜色最深;元值 $h$ 出现在2个关联规则中,因此颜色次之;元值 $c, d, f, j, l, m, o$ 出现在1个关联规则中,因此颜色最浅;其它元值未出现在某条关联规则中,因此呈透明色。依据前注意加工理论,一些视觉特征(如:颜色、线条方向、线条长度和宽度、闭环等)可以被低层视觉系统快速、精确地感知。读者可以容易地识别出重要元值,进而判断该元值属于哪个元属性。

#### 4 算法实现

本文数据来源于某省全员人口数据库,以育龄妇女为核心的夫妇年龄差及其影响因素进行关联规则挖掘结果的可视化。心理学家用纺锤体在三维空间中表示颜色的基本特征,称为颜色纺锤体。该颜色模型的垂直轴线表示白黑系列的亮度变化,水平圆周表示不同的色调,半径方向表示饱和度的变化,纺锤体是形似纺锤的结构。

本文借鉴颜色纺锤体的思想,设计符合关联规则可视化任务的纺锤体结构。基于S-C元图结合纺锤体的关联规则可视化表示方法的算法思想:首先将关联规则挖掘中的元值等间距排列成椭圆,椭圆中心为屏幕中心,并画出纺锤体中心

轴;然后用元边将关联规则前件与后件中的所有元值节点连接起来;接着为元值着色,以体现重要性;最后输出以S-C元图结合纺锤体展示关联规则的结果。具体实现过程如下。

##### 算法1 SetRule()//关联规则的S-C元图结合纺锤体展示

输入:按置信度从小到大排列的关联规则集合 RuleSets, 包括 AR\_ID, LHS, RHS, Supp, Conf; 椭圆中心坐标  $O(x_0, y_0)$

相邻两条元规则P点的间隔  $h$

输出:关联规则的S-C元图结合纺锤体表示 S-C\_Spindle

SetRule(RuleSets)

```

1) SetEllipse(K); //调用算法2
2) DrawCenterLine( $x_0, y_0$ ); //过椭圆中心画垂直的双向中心轴
3) FOR Rs  $\in$  RuleSets DO { //画规则集中的每条元规则
4)  $X_{gi} = X_0, Y_{gi} = Y_0 + (-1)^i * h * \lceil i/2 \rceil$ ; //存放支持度和置信度的P点的坐标  $g_i$ 。第奇数个规则画在椭圆平面上部;第偶数个规则画在椭圆平面下部。
5)  $r_i = |Supp_i|/2$ ; //P的半径
6) DrawCircle( $g_i, r_i, color_i$ ); //画存放支持度和置信度的点P
7) IF (AR_ID == odd) THEN { //第奇数个规则画在椭圆平面上部
8)   FOREACH  $k_j \in$  LHS DO { //规则前件中的每个元值
9)     IF ( $y_{LHSj} < y_0$ ) THEN { //如果规则前件元值  $k_j$  的纵坐标小于  $y_0$ 
10)      DrawLine( $p_{j2}, g_i, color_i$ ); //colori 随 i 的增大而变深,为暖色系
11)    ELSE
12)      DrawLine( $p_{j1}, g_i, color_i$ );
13)    } //end foreach
14)   FOREACH  $k_f \in$  RHS DO { //规则后件中的每个元值
15)     Draw directed line in the same way that  $k_j$  did;
16)   } //end foreach
17) } ELSE { //第偶数个规则画在椭圆平面下部
18)   FOREACH  $k_w \in$  LHS DO { //规则前件中的每个元值
19)     IF ( $y_{LHS w_j} < y_0$ ) THEN { //如果规则前件元值  $k_w$  的纵坐标小于  $y_0$ 
20)      DrawLine( $p_{w1}, g_i, color_i$ ); //colori 随 i 的增大而变深,为冷色系
21)    ELSE
22)      DrawLine( $p_{w2}, g_i, color_i$ );
23)    } //end foreach
24)   FOREACH  $k_s \in$  RHS DO { //规则后件中的每个元值
25)     Draw directed line in the same way that  $k_w$  did;
26)   } //end foreach
27) } //end for
28) Cout[] = CountEach( $\sum k_s + \sum k_f$ ); //求每个元值出现的次数
29) SetColorToCout[e] from max to min; //依据 Cout[e] 从大到小为元值着色,灰度值依次降低

```

##### 算法2 SetEllipse()//将元值库 $K = \{k_i, i=1, 2, \dots, I\}$ 中的

的元值布局为以  $O(x_0, y_0)$  为中心的椭圆

输入:元值库  $K$ ; 椭圆中心坐标  $O(x_0, y_0)$ ; 椭圆长半轴  $m$ ; 椭圆短半轴  $n$ ; 元值半径  $r$

输出:元值  $\{k_i, i=1, 2, \dots, I\}$  的布局  $K\_Layout$ ; 元值中心和基准椭圆中心连线与C的内侧和外侧交点坐标  $p_{i1}, p_{i2}$

SetEllipse(K)

```

1) BaseEllipse = DrawEllipse( $O, m, n$ ); //画元值中心点所在的基准椭圆
2) BaseEllipse, border = None; //将基准椭圆隐藏
3) AssistEllipse1 = DrawEllipse( $O, m-2*r, n-2*r$ ); //画基准椭圆内侧辅助椭圆

```

4) AssistEllipse2=DrawEllipse(O, m+2 \* r, n+2 \* r); //画基准椭圆  
 外侧辅助椭圆 AssistEllipse1 和 AssistEllipse2 在构建元属性 A 时使用  
 5)  $X_{i1}=X_0, Y_{i1}=Y_0-n$ ; //第一个元值的中心位置  
 6) FOR  $k_i \in K$  DO { //遍历元值库 K 中的每个元值, 构建椭圆  
 7)  $C=\text{DrawCircle}(t_i, r)$ ; //画代表第 i 个元值的圆  
 8)  $\text{Item}(C)_i = k_i, \text{item}$ ; //将 C 的文本置为第 i 个元值的值  
 9)  $\text{Compute}(p_{i1}, p_{i2})$ ; //  $p_{i1}, p_{i2}$  分别为元值中心和基准椭圆中心连线与 C 的内侧和外侧交点  
 10) IF ( $k_i = \text{First}(A_{\text{current}})$ ) THEN { //如果  $k_i$  是当前元属性的第一个元值, 则逆时针画弧线连接  $q_{i1}, q_{i2}$   
 11)  $\text{LinkArc}(q_{i1}, q_{i2})$ ;  
 12) }  
 13) IF ( $k_i = \text{Last}(A_{\text{current}})$ ) THEN { //如果  $k_i$  是当前元属性的最后一个元值, 则逆时针画弧线连接  $q_{i2}, q_{i1}$   
 14)  $\text{LinkArc}(q_{i2}, q_{i1})$ ;  
 15) }  
 //  $q_{i1}, q_{i2}$  分别为元值中心和基准椭圆中心连线与内侧辅助椭圆和外侧辅助椭圆的交点  
 16)  $j=i$ ;  
 17)  $i++$ ;  
 18)  $t_i = \text{next}(t_{i-1}, \Delta l)$ ; //计算下一个元值的坐标  
 19) Delete redundant curve;  
 20) }  
 21) //end for  
 22) return K\_Layout,  $p_{i1}, p_{i2}$ ;

上述两个算法所用公式如下:

基准椭圆周长:  $l = 2 * \pi * n + 4 * (m - n)$ ,  $\Delta l = l / (N - 1)$ , 其中  $N$  是元值个数。

过椭圆中心与元值中心的直线 B 方程:

$$y = (y_0 - y_{i1}) * (x - x_0) / (x_0 - x_{i1}) + y_0 \quad (1)$$

内侧辅助椭圆 AssistEllipse1 的方程:

$$(x - x_0)^2 / (m - 2 * r)^2 + (y - y_0)^2 / (n - 2 * r)^2 = 1 \quad (2)$$

外侧辅助椭圆 AssistEllipse2 的方程:

$$(x - x_0)^2 / (m + 2 * r)^2 + (y - y_0)^2 / (n + 2 * r)^2 = 1 \quad (3)$$

代表元值的圆 C 方程:

$$(x - x_{i1})^2 + (y - y_{i1})^2 = r^2 \quad (4)$$

由式(1)、式(2)求得元值中心和基准椭圆中心连线与内侧辅助椭圆交点中距元值中心较近的点  $q_{i1}$ 。

由式(1)、式(3)求得元值中心和基准椭圆中心连线与外侧辅助椭圆交点中距元值中心较近的点  $q_{i2}$ 。

由式(1)、式(4)求得元值中心和基准椭圆中心连线与 C 的内侧和外侧交点  $p_{i1}, p_{i2}$ 。

为提高视觉区分度, 算法 1 和算法 2 将展示上部由中心到外侧颜色逐渐变深的暖色系圆锥面, 下部由中心到两侧颜色逐渐变深的冷色系圆锥面, 最终构成对称的纺锤体形状。每条元边颜色不同, 箭头指向关联规则后件元值, 同时保存支持度和置信度信息的 P 点以大小和位置向用户提供信息, 体现了前注意加工理论。用户不需要耗费过多的注意资源, 便能对环境中的重要信息自动处理。每条元边的颜色在整体中是独一无二的, 易被感知为整体, 体现了格式塔理论。用户根据元边的布局, 利用经验将每条元边感知为一个锥面。

本文算法采用某省全员人口数据库中的育龄妇女信息记录, 其中 A、B、C、D、E、F 代表所选信息记录属性。本节仅对某条规则 001  $\{E = e_3\} \Rightarrow \{A = a_2, C = c_1\}$   $\{Supp = 30.4\%,$

$Conf = 15.7\%\}$  进行展示。其中, 按置信度从小到大排列, AR\_ID(001) 代表关联规则序号, LHS(left-hand-side) 代表规则前件, RHS(right-hand-side) 代表规则后件, Supp 代表支持度, Conf 代表置信度。根据算法 1 第 1)、2) 步, 以椭圆形式展示元属性 A、B、C、D、E、F 及其元值, 并画出中心轴, 如图 7 所示。

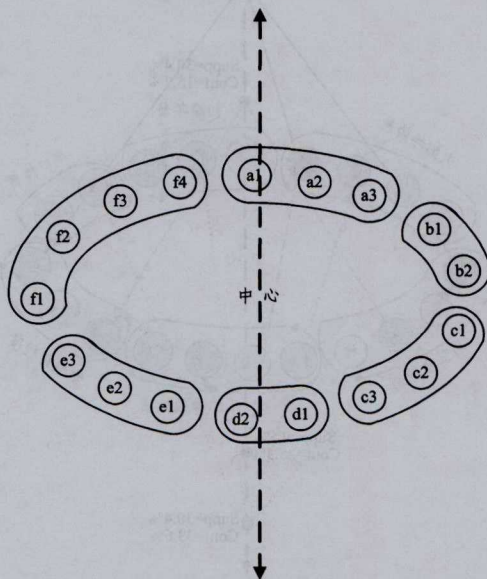


图 7 依据算法 1 和算法 2 布局元值

规则 001 是第奇数个规则, 因此依据算法 1 第 4) — 19) 步, 将元边绘制在椭圆平面上方。由于是第一个规则, 保存支持度和置信度的点 P 的中心在椭圆中心上方  $h$  处。为避免在平面上元边将元值遮住, 判断规则 001 的前件和后件中元值中心坐标的位置。前件  $e_3$  纵坐标小于椭圆中心点纵坐标, 因此连接过 P 点中心和直线 B 与元值内侧的交点  $p_{e31}$ , 同样判断后件  $a_2$  和  $c_1$  的位置链接相应线段。因置信度最低, 所以选择暖色系中较浅色。依据算法 1 第 34)、35) 步, 将 3 个元值置为浅灰色, 其余元值为透明色, 如图 8 所示。

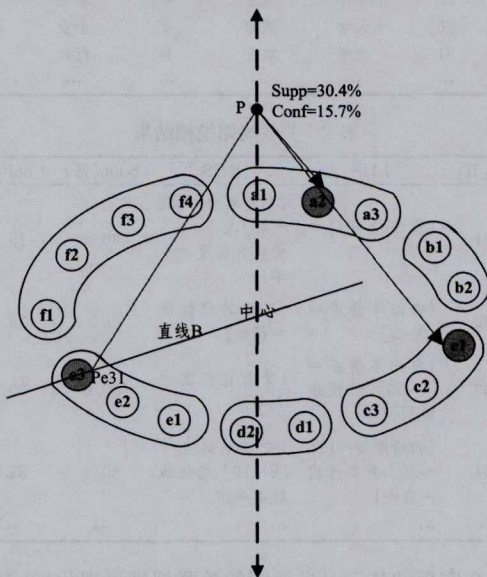


图 8 依据算法 1 和算法 2 布局关联规则

## 5 实验结果与分析

本文以某省全员人口数据库为数据源对算法进行了实际

测试和验证。选取6个以育龄妇女为核心的夫妇年龄差及其可能影响因素:初婚年龄、受教育程度、户口性质、居住地级别、工作单位性质、夫妇年龄差,对关联规则挖掘结果进行可视化,如图9所示。

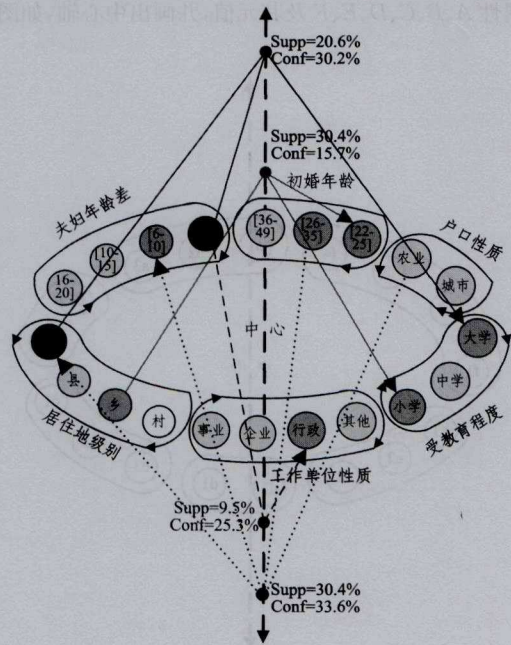


图9 人口数据的关联规则展示

表1列出了育龄妇女信息记录,ID代表育龄妇女人口编码。表2列出了由表1的属性集合所挖掘出的关联规则结果,按置信度从小到大排列,AR\_ID代表关联规则序号,LHS代表规则前件,RHS代表规则后件,Supp代表支持度,Conf代表置信度。

表1 育龄妇女信息记录表

| ID   | 初婚年龄 | 受教育程度 | 户口性质 | 居住地级别 | 工作单位性质 | 夫妇年龄差 |
|------|------|-------|------|-------|--------|-------|
| 0001 | 23   | 小学    | 农业   | 村     | 其他     | 1     |
| 0002 | 25   | 中学    | 农业   | 乡     | 事业     | 5     |
| 0003 | 36   | 大学    | 城市   | 市     | 企业     | 3     |
| 0004 | 31   | 大学    | 农业   | 县     | 行政     | 7     |
| ...  | ...  | ...   | ...  | ...   | ...    | ...   |

表2 关联规则挖掘结果

| AR_ID | LHS                    | RHS                     | Supp(%) | Conf(%) |
|-------|------------------------|-------------------------|---------|---------|
| 001   | {居住地级别=乡}              | {初婚年龄=[22~25],受教育程度=小学} | 30.4    | 15.7    |
| 002   | {夫妇年龄差=[0~5]}          | {工作单位性质=行政}             | 9.5     | 25.3    |
| 003   | {夫妇年龄差=[0~5],居住地级别=市}  | {受教育程度=大学}              | 20.6    | 30.2    |
| 004   | {初婚年龄=[26~35],户口性质=农业} | {夫妇年龄差=[6~10],居住地级别=市}  | 30.4    | 33.6    |
| ...   | ...                    | ...                     | ...     | ...     |

本文实现的基于S-C元图的关联规则可视化形式与传统文献相比具有以下优点:

有效地克服了传统关联规则可视化方法存在的无法进行多模式展示<sup>[2]</sup>、规则属性增多时边或节点重叠<sup>[1-6]</sup>、不能展示支持度和置信度<sup>[5]</sup>、不适合普通用户理解<sup>[4]</sup>、无法展示规则总

体特征<sup>[3]</sup>等问题;实现了多模式关联规则可视化展示,通过S-C元图宏观呈现元属性与元值的关系;规则前后件信息区分明显,规则之间不会重叠;借助于纺锤体使每条元边各自形成一个锥面,避免了元边的交叉;运用前注意加工理论和格式塔理论,考虑普通用户的认知特点,规则的可解释性较好。

**结束语** 本文首次提出S-C元图的概念及其应用于关联规则表示的相关定义,设计了一种新的基于S-C元图结合纺锤体的关联规则可视化算法,并应用于某省全员人口数据分析和应用。该方法运用前注意加工理论和格式塔理论将元规则所涉及到的元属性和元值关系展示给用户,且能展示元规则的信息。实验结果表明,该方法避免了多条关联规则元边的交叉,更有助于用户对关联规则的整体感知,便于用户直观地识别重要元值或元属性,减轻了用户的感知负担,提高了用户决策效率。在下一步的研究中,我们将针对元值较多的元属性,根据属性的相对重要程度,对算法的交互性进行改进。

参考文献

[1] Buono P, Costabile M F. Visualizing association rules in a framework for visual data mining [C]// From Integrated Publication and Information Systems to Information and Knowledge Environments. Springer Berlin Heidelberg, 2005: 221-231

[2] Ertek G, Demiriz A. A framework for visualizing association mining results [C]// Proceedings of the 21st international conference on Computer and Information Sciences (ISCIS'06). Berlin: Springer-Verlag, 2006: 593-602

[3] Hahsler M, Chelluboina S. Visualizing Association Rules in Hierarchical Groups [C]// 42nd Symposium on the Interface: Statistical, Machine Learning, and Visualization Algorithms (Interface 2011). America, 2011: 1-11

[4] Liu G, Suchitra A, Zhang H, et al. AssocExplorer: an association rule visualization system for exploratory data analysis [C]// Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. New York: ACM, 2012: 1536-1539

[5] Bornelöv S, Enroth S, Komorowski J. Visualization of Rules in Rule-Based Classifiers [C]// Intelligent Decision Technologies Smart Innovation, Systems and Technologies (IDT'2012). Berlin: Springer-Verlag, 2012: 15: 329-338

[6] Cinicioglu E N, Ertek G, Demiriz D, et al. A framework for automated association mining over multiple databases [C]// Proceedings of International Symposium on Innovations in Intelligent Systems and Applications. Istanbul, Turkey: IEEE, 2011: 79-85

[7] 郭晓波, 赵书良, 刘军丹, 等. 基于概念图的关联规则知识表示 [J]. 计算机科学, 2013, 40(8): 261-265

[8] Bruzzese D, Davino C. Visual mining of association rules [C]// Visual Data Mining: Theory, Techniques and Tools for Visual Analytics (LNAI 6208). Berlin: Springer-Verlag, 2008: 103-122

[9] Hahsler M, Chelluboina S. Visualizing Association Rules: Introduction to the R-extension Package arulesViz [J]. R project module, 2011: 223-238

[10] Toker D, Conati C, Carenini G, et al. Towards adaptive information visualization; on the influence of user characteristics [C]// User Modeling, Adaptation, and Personalization. Berlin: Spring-

- [11] Steichen B, Carenini G, Conati C. User-adaptive information visualization; using eye gaze data to infer visualization tasks and user cognitive abilities[C]// Proceedings of the 2013 international conference on intelligent user interfaces. New York: ACM, 2013; 317-328
- [12] Fekete J D, Van Wijk J J, Stasko J T, et al. The value of information visualization [C]// Information Visualization. Berlin: Springer-Verlag, 2008; 1-18
- [13] Peterson D J, Berryhill M E. The Gestalt principle of similarity benefits visual working memory [J]. Psychonomic bulletin & review, 2013; 1-8
- [14] Rusu A, Fabian A J, Jianu R. Using the gestalt principle of closure to alleviate the edge crossing problem in graph drawings [C]// Proceedings of the 2011 15th International Conference on Information Visualisation (IV). London: IEEE, 2011; 488-493

(上接第 266 页)

下,本文方法的相似度计算精度明显优于其他两种方法,在阈值较大时相似度计算精度的差异变小。分析表 2 和表 4,本文提出的方法较文献[11]的相似度计算精度平均提高了 5% 和 7%,较文献[10]平均提高了 8% 和 13%。

(2)在属性缺少情况频繁出现的数据集中,例如 FOAF 数据集,在实例相似度计算时,foaf-name 以及 foaf-topic 等重要属性都存在缺失情况,在计算时会导致权值的改变。本文方法计算的相似度精度明显优于其他两种方法的计算精度,究其原因,表 2 中,本文方法的优越性主要在于考虑了属性的取值类型影响因素,表 4 中,本文方法的优越性不仅与属性的取值类型有关,还与属性的缺失情况有关。从这两个表中,也可以得出属性的取值类型以及属性的缺失情况对实例相似度计算都有影响。

(3)分析表 3 和表 5,本文提出的方法相对于其他两种方法较稳定。

**结束语** 针对当前传统的关联数据语义相似度计算方法考虑因素单一,没有充分利用实例属性的语义信息而造成相似度计算结果精确性不高的问题,本文提出一种基于动态权值的关联数据语义相似度计算方法。该方法首先根据实例属性不同取值的数量、属性值的分布以及属性的有效性 3 个因素对实例属性的权值进行动态确定,然后根据属性类型的不同自动选择不同的相似度计算方法。实验表明,本文提出的相似度计算方法充分利用了实例属性的语义信息,能更准确表示出实例之间的语义相似关系,提高了计算的精度。但是本文提出的算法针对的只是单个数据集,没有考虑跨数据集实例对的相似度计算以及计算精度与阈值的关系,而且计算性能也待进一步提高以应对大规模数据集的情况。下一步拟在本文的基础上,进一步研究跨数据集的实例对的语义相似度计算,并将新计算方法应用于关联数据应用领域,以提高检索效率。

## 参 考 文 献

- [1] Berners-Lee T. Linked data-the story so far [J]. International

- [15] Healey C G, Enns J T. Attention and visual memory in visualization and computer graphics [J]. IEEE Transactions on Visualization and Computer Graphics, 2012, 18(7): 1170-1188
- [16] Michalski R, Grobelny J. The role of colour preattentive processing in human-computer interaction task efficiency: a preliminary study [J]. International Journal of Industrial Ergonomics, 2008, 38(3): 321-332
- [17] Cui W, Qu H. A survey on graph visualization [D]. Hong Kong University of Science and Technology, 2007; 16-29
- [18] Basu A, Blanning R. Metagraphs and Their Applications [M]. Berlin: Springer-Verlag, 2006; 1-11, 77-115
- [19] 谭政华, 胡光锐, 任晓林. 模糊元图及其特性分析[J]. 计算机研究与发展, 2000(3): 272-277
- [20] Dashore P, Jain S, Dashore S R. Fuzzy Metagraph and Rule Based System for Decision Making in Share Market [J]. International Journal of Computer Applications, 2010, 6(2): 10-13

Journal on Semantic Web and Information Systems, 2009, 5(3): 1-22

- [2] Servant F P. Linking enterprise data [C]// Linked Data on the Web. 2008
- [3] Hartig O, Sequeda J, Taylor J, et al. How to consume Linked Data on the Web; tutorial description[C]// Proceedings of the 19th international conference on World Wide Web. ACM, 2010; 1347-1348
- [4] Tversky A. Features of similarity [J]. Psychological review, 1977, 84(4): 327-352
- [5] 高学东, 吴玲玉, 武森, 等. 基于属性与对象关系信息的综合差异度计算[J]. 计算机工程, 2011, 37(22): 35-38
- [6] Sheth A, Aleman-Meza B, Arpinar I B, et al. Semantic association identification and knowledge discovery for national security applications[J]. Journal of Database Management, 2005, 16(1): 33-53
- [7] 刘宏哲, 须德. 基于本体的语义相似度和相关度计算研究综述[J]. 计算机科学, 2012, 39(2): 8-13
- [8] Bhattacharya I, Getoor L. Iterative record linkage for cleaning and integration [C]// Proceedings of the 9th ACM SIGMOD Workshop on Research Issues in data Mining and Knowledge Discovery. ACM, 2004; 11-18
- [9] Zadeh P D H, Reformat M Z. Fuzzy semantic similarity in linked data using the OWA operator[C]// Fuzzy Information Processing Society (NAFIPS). 2012 Annual Meeting of the North American. IEEE, 2012; 1-6
- [10] Song D, Heflin J. Domain-independent entity reference in RDF graphs[C]// Proceedings of the 19th ACM International Conference on Information and Knowledge Management. ACM, 2010; 1821-1824
- [11] 张晓辉, 蒋海华, 邸瑞华. 基于属性权重的链接数据共指关系构建[J]. 计算机科学, 2013, 40(2): 40-43
- [12] Glaser H, Millard I C, Jaffri A. RKBExplorer. com: a knowledge driven infrastructure for linked data providers[M]// The Semantic Web: Research and Applications. Springer Berlin Heidelberg, 2008; 797-801