探索关联规则可视化的结构化关联映射图

易 黎[1] 胡雅萌[1,2] 彭艳兵[1]

[1](南京烽火软件科技有限公司 江苏 南京 210019)

[2](武汉邮电科学研究院 湖北 武汉 430074)

摘 要 对于大量的高维度的交易数据,利用关联规则进行数据挖掘,用户难以进行解释和利用。主要两个原因:常规关联规则挖掘算法可产生大量关联规则;一些关联规则可部分重叠。若用户能自主选择,在关联规则挖掘中所使用的相关项集,则可解决该问题。提出一种新的视觉探索工具,结构化关联映射图,使用户能够以视觉方式找到相关项集的组。该方法使用健康检查结果数据集进行验证,并且实验结果表明具有最高 2×2 规则贡献的和值的结构化关联映射图有助于显著减少关联分析的复杂性,并且能够集中于搜索空间的特定区域关联规则挖掘,同时避免不相关的关联规则。

关键词 可视化 关联规则挖掘 分层聚类 结构化关联映射图

中图分类号 TP391.4 **文献标识码** A DOI:10.3969/j.issn.1000-386x.2017.12.013

EXPLORING STRUCTURED ASSOCIATION MAP OF ASSOCIATION RULES VISUALIZATION

Yi Li[1] Hu Yameng[1,2] Peng Yanbing[1]

[1](*Fiber Home Starry Sky Co.,Ltd.,Nanjing 210019,Jiangsu,China*)

[2](*Wuhan Research Institute of Posts and Telecommunications,Wuhan 430074,Hubei,China*)

Abstract The users often face difficulties in interpreting and exploiting the association rules extracted from large transaction data with high dimensionality. There are two main reasons. Firstly, too many association rules can be produced by the conventional association rule mining algorithms, and secondly, some association rules can be partly overlapped. This problem can be solved if the users can select the relevant items to be used in association rule mining. In this context, this paper aims to propose a new visual exploration tool, structured association map, which enables the users to find the group of the relevant items in a visual way. For illustration, this procedure is applied to a mass health examination result data set, and the experiment results demonstrate that structured association map with maximum sums of 2×2 regular contributions value helps to reduce the complexities of association analysis significantly and it enables to focus on the specific region of the search space of association rule mining while avoiding the irrelevant association rules.

Keywords Visualization Association rule mining Hierarchical clustering Structured association mapping

0 引 言

随着生活水平的提高,预防保健成为公众关注的焦点,Boulware 等[1]认为大众健康检查(MHE)在监测和评估个人健康水平方面发挥了重要作用,Kweon 等[2]也提到 MHE 结果数据为在国家和个人层面制定卫生政策或战略提供了坚实的基础。边根庆等[3]表明了数据挖掘能为我们提供有价值的重要信息或知识,从而产生不可估计的经济效益。李春青[4]认为关联规则算法是数据挖掘算法中重要的分析方法,能够挖掘数据中各项关联。然而,分析从大众健康检查收集来的数据集相当困难,因为它们包括许多相关变量,探索高维交易数据内的关联规则时,数据过于复杂抽象,因而难以被直观的展示出来。

肖晗等[5]提出通过数据挖掘产生的关联规则会存

收稿日期:2017-05-15。**易黎**,工程师,主研领域:大数据分析。**胡雅萌**,硕士生。**彭艳兵**,教授级高级工程师。

在大量无用和不感兴趣的规则,同时刘晓蔚[6]提到传统的类关联规则挖掘算法在挖掘完整的规则数据集时往往需要消耗很长的时间。此外,Yang[7]提到关联规则的前提和后果是在所有项的集合的幂集上定义的,并且它们项之间表现出了多对多的关系。而 Ferreira 等[8]提出可视化技术可以处理大量而复杂的规则。

用于表示大量关联规则的最常见和简单的方法是表格。由于其简单性,基于表格的视图用于许多常规数据挖掘软件中,并且这种表格中的规则通常通过诸如置信度或提升的兴趣度度量排序。然而,Sekhavat 等[9]提出若发现太多关联规则,分析器在解释列表和从表中找到有趣的规则仍有困难。

本文介绍了一种称为结构化关联图的新型可视化方法,是关联规则集合簇热图的变体,用于总结高维交易数据中二元变量之间的关系。所提出的方法基于关联规则挖掘和聚类分析的常规数据挖掘技术,并且其使得用户能够容易地找到由一组相关联的二进制变量形成的感兴趣区域,这个区域可构成感兴趣的许多关联规则。由于结构化关联映射图是基于矩阵的方法,它很容易实现和解释。与基于经典矩阵的技术相比,结构化关联映射图更适合于解释给定项集间的多对多关系。

1　研究方法

1.1　挖掘过程

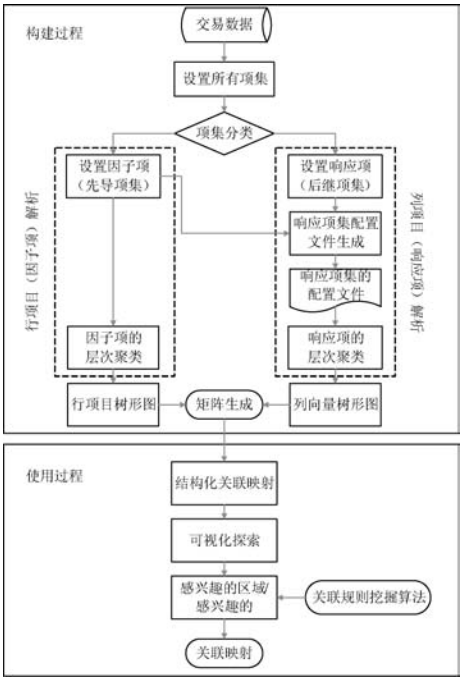

图 1　基于结构化关联映射图的关联规则挖掘过程

图 1 描述了基于结构化关联映射图的关联规则挖掘过程,上半部分描述其构建阶段,下半部分总结其利用阶段。在构建阶段,矩阵与以不同方式构造的两个树形图组合,获得结构化关联映射图。创建后,结构化关联映射图用于可视化识别感兴趣的区域和组,感兴趣组由关联规则挖掘算法探索。

1.2　行项集(因素项)分析

行项集分析的目的是通过对因子项应用层次聚类算法来生成因子项树形图。Michael 等[10]提出式(1)中的亲和度是两个项集 a 和 b 的相似性度量,而式(2)中的 Jaccard 距离可以用于测量它们之间的距离,其中 $\sup(X)$ 表示项集集合 X。在本文中,式(2)中的距离度量用于生成因子项集树形图。

$$A(a,b) = \frac{\sup(\{a,b\})}{\sup(\{a\}) + \sup(\{b\}) - \sup(\{a,b\})} \quad (1)$$

$$J_d(a,b) = 1 - A(a,b) \quad (2)$$

$\boldsymbol{D}_F$ 因子项的平方距离矩阵可如式(3)获得,其中 m_f 是因子项的数量,并且 $df_{ij} = J_d(F_i,F_j)$,F_i 和 F_j 为因子项集。注意,如果 $i = j$ 且 $df_{ij} = df_{ji}$,则 $df_{ij} = 0$。

$$\boldsymbol{D}_F = \begin{bmatrix} df_{11} & df_{12} & \cdots & df_{1m_f} \\ df_{21} & df_{22} & \cdots & df_{2m_f} \\ \vdots & \vdots & & \vdots \\ df_{m_f1} & df_{m_f2} & \cdots & df_{m_fm_f} \end{bmatrix} \quad (3)$$

在本文中,凝聚层次聚类算法应用于距离矩阵 $\boldsymbol{D}_F$,以生成树形图。聚集聚类算法需要确定两个聚类(项集)之间的距离的链接标准。Tan 等[11]提出常用的链接标准是单链(SL)、完全链(CL)、平均链(AL)和 Ward's 标准(WC)。本文通过对比四种标准,找出最优值。而用于对树形图的子树进行排序的方法使用基于支持度量(OM),这是一个简单的自上而下排序方法,从最高合并点开始。在每个合并点,此方法查找哪个子树具有支持最高的项集,并将其放在树形图的左侧(上侧)。

1.3　列项(响应项)分析

通过层次聚类算法获得响应项 $\boldsymbol{D}_R$ 的距离矩阵,来生成响应项树形图。

在本文中,响应项 $\boldsymbol{R}_j$ 的定义如下:

$$PF(R_j) = [L_{1j}, L_{2j}, \cdots, L_{m_fj}] \quad (4)$$

式中:L_{ij} 是 F_i 对 R_j 的影响。规则"$\{F_i\} \rightarrow \{R_j\}$"的兴趣度量用 L_{ij} 表示,并且本文通过使用升力测量来计算 L_{ij} 如下:

$$LIFT(\{F_i\} \rightarrow \{R_j\}) = \frac{\mathrm{conf}(\{F_i\} \rightarrow \{R_j\})}{\sup(\{R_j\})} \quad (5)$$

考虑因子项的影响的分布,R_j 和 R_k 之间的距离

dr_{jk} 计算如下：

$$dr_{jk} = 1 - \frac{PF(R_j) \cdot PF(R_k)}{|PF(R_j)| \times |PF(R_k)|} \quad (6)$$

式中：$PF(R_j) \cdot PF(R_k)$ 是两个轮廓向量 $PF(R_j)$ 和 $PF(R_k)$ 的内积，$|PF(R_j)|$ 是 $PF(R_j)$ 的长度。

1.4 结构化关联映射图的评价

本文引用一种基于兴趣的评估方法，即 2×2 规则贡献的和，由“相邻”项组成的概念。

如下计算 2×2 规则贡献的和测量：

$$S2C = \sum_{i=1}^{m_f-1} \sum_{j=1}^{m_r-1} CN(\{F_{(i)}, F_{(i+1)}\} \rightarrow \{R_{(j)}, R_{(j+1)}\}) \quad (7)$$

式中：$CN(\{F_{(i)}, F_{(i+1)}\} \rightarrow \{R_{(j)}, R_{(j+1)}\})$ 是 $\{F_{(i)}, F_{(i+1)}\} \rightarrow \{R_{(j)}, R_{(j+1)}\}$ 的规则。如果先导和后继项集都连接到缩减的结构化关联映射图的树形图中，则该规则被关闭，并且其贡献计算如下：

$$CN(\text{closedrule}) = LIFT(\text{closedrule}) \quad (8)$$

式(8)中的 CN(closedrule)表示闭合规则应当具有高兴趣度值，因此分析器倾向于期望它们是有趣的规则。

如果在缩减结构化关联映射图的树形图中既没有连接先导项也没有连接后继项，则打开关联规则，并且如下获得其贡献：

$$CN(\text{openedrule}) = \begin{cases} LIFT(\text{openedrule})^{-1} & \text{if } LIFT(\text{openedrule}) > 0 \\ M & \text{otherwise} \end{cases} \quad (9)$$

式(9)中的 CN(openedrule)意味着打开的规则具有低兴趣度值。因此，打开的规则的贡献是其升力的倒数，如果它具有正升力值。如果打开规则的提升为0，则规则的贡献被设置为任意值 M，并且本文中 $M = 1$。

1.5 结构化关联映射图利用率

结构化关联映射图用于可视化探索给定事务数据内的二进制变量之间的关系。并且，结构化关联映射图通过应用关联规则挖掘算法帮助用户更深入地找到要研究的感兴趣区域。

设 $S(a,b,p,q)$ 表示由代表 $F_{(a)}, F_{(a+1)}, \cdots, F_{(a+p-1)}$ 的行和代表 $R_{(b)}$、$R_{(b+1)}$、$R_{(b+q-1)}$ 的列，当 $p, q \geqslant 2$ 时，组成的结构化关联映射图的 $p \times q$ 子矩阵。如果满足以下两个条件，则将 $S(a,b,p,q)$ 称为感兴趣区域：

(1) $S(a,b,p,q)$ 中的几乎所有 $e_{ij}s$ 指示 $F_{(i)}$ 和 $R_{(j)}$ 之间的正相关，或 $e_{ij}s$ 指示 $F_{(i)}$ 和 $R_{(j)}$ 之间的负相关 $(a \leqslant i \leqslant a+p-1, b \leqslant j \leqslant b+q-1)$。

(2) $F_{(a)}, F_{(a+1)}, \cdots, F_{(a+p-1)}$ 在要素项集树形图中以较低的级别合并，且 $R_{(b)}$、$R_{(b+1)}$、$R_{(b+q-1)}$ 在响应项树形图中的较低级合并。

如果 $S(a,b,p,q)$ 是感兴趣区域，则与该子矩阵 $G(a,b,p,q)$ 相关的项集合被称为感兴趣组。

$$G(a,b,p,q) = \{F_{(a)}, F_{(a+1)}, \cdots, F_{(a+p-1)}\} \cup \{R_{(b)}, R_{(b+1)}, \cdots, R_{(b+q-1)}\} \quad (10)$$

2 应用实例

2.1 大众健康检查结果数据集

原始数据从韩国 278 个青少年的大众健康检查收集。其中，两类变量表示受试者的身体状况或病史，由专职医务人员检查。其他类别中的变量表示主观症状和个人生活方式，基于受试者对感知健康状况的陈述。图 2 为大众健康检查结果数据集的项集分类。

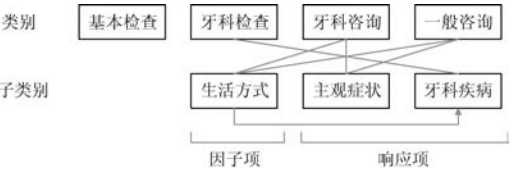

图 2 大众健康检查结果数据集的项集分类

通过使用结构化关联映射图来探寻个人牙齿健康可视化变量之间的关系。本文选择生活方式变量作为因子项，主观症状和牙科疾病变量作为响应项，分析形式“$\{I_{\text{life}}\} \rightarrow \{I_{\text{symptom}} \cup I_{\text{disease}}\}$”的关联规则。

2.2 结 论

一旦项集被分类，下一步是生成因子项集树形图。为此，凝聚层次聚类算法应用于因子项的距离矩阵 $\boldsymbol{D}_F$，通过使用式(2)获得。

然后，我们可以通过使用式(4)－式(6)生成响应项，并且这些响应项用于计算响应项 $\boldsymbol{D}_R$ 的距离矩阵。同样，响应项树形图通过应用层次聚类算法获得。

由于我们有 4 个因素项集树状图和相同数量的响应项集树状图，可以构造 4 个不同的结构化关联映射图。可以通过使用式(8)－式(10)中描述的 2×2 规则贡献的和测量来评价它们的性能，评价结果总结在表 1 中。在表 1 中，每行指示因子项集树形图的排序方法和链接标准，而每一列指定响应项集树形图的排序方法和链接标准。

表 1 不同结构化关联映射的 2×2 规则贡献的和值

行＼列	AL	SL	CL	WC
AL	41.3	38.19	37.14	37.14
SL	41.3	38.19	37.14	37.14
CL	41.4	37.17	36.97	36.97
WC	51.41	49.24	45.22	45.22

(下转第 157 页)

[4] 冯明月,章永龙,浦宏艺. 基于 B/S 模式的 O2O 汽车服务平台的设计与实现[J]. 计算机应用与软件,2017,34(6):57-61.
[5] 胡颖辉,宁赛飞. 基于 UML 和 Asp. Net 实现三层 B/S 结构系统开发[J]. 计算机与信息技术,2007(Z1):43-50.
[6] 卢小晨,曲震,马泽方,等. 论互联网+大数据算法在税收工作中的应用[J]. 税务研究,2017,2(1):14-16.
[7] 赵建英. 基于. NET 三层架构的旅游电子商务平台的设计与开发[D]. 上海:复旦大学,2013.
[8] 马超,徐迭实,张淑丽,等. 大数据环境下离散制造车间异常事件发现方法[J]. 计算机应用与软件,2017,34(9):288-293.
[9] 王晓燕,陈晋川,杜小勇. 云计算环境中面向 OLTP 应用的数据分布研究[J]. 计算机学报,2016,39(2):253-269.
[10] 袁弘,张明江,李建祥,等. 基于云服务的电动汽车电池安全预警系统设计[J]. 计算机应用与软件,2014,31(9):63-66.

(上接第 70 页)

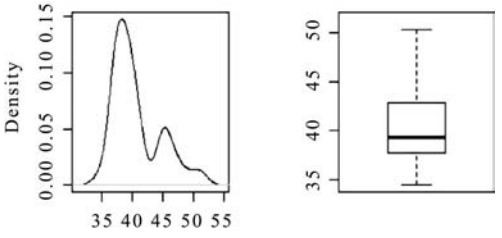

图 3　2×2 规则贡献的和值的分布

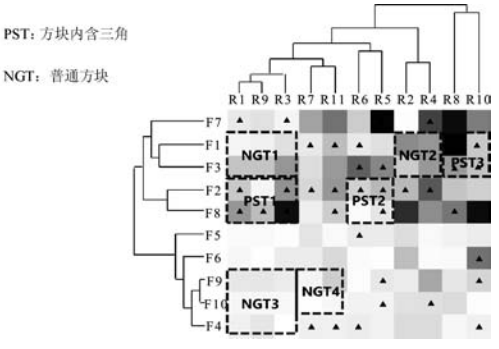

图 4　结构化关联映射图与最高 2×2 规则贡献的和
(结构化关联映射图与 OM-WC\OM-AL)

表 1 的第 i 行和第 j 列中的元素表示通过组合第 i 个因子项集树形图和第 j 个响应项集树形图构建的结构化关联映射图的 2×2 规则贡献的和值,并且可以看出结构化关联映射图的性能根据组合树形图的结构而显着变化,如图 3 所示。在表 1 中列出的 16 个不同的结构化关联映射图中,我们必须选择具有 WC/AL 的结构化关联映射图(因子项集树形图结构/响应项集树状图结构),因为它使 2×2 规则贡献的和测量的值最大化。因此,我们可以获得如图 4 所示的优化的结构化关联映射图。其中每个方块 e_{ij} 根据关联规则"$\{F_{(i)}\} \rightarrow \{R_{(j)}\}$"的提升值被着色,如下:① 方块内含三角指示升力值高于 1(正相关),而普通方块意味着升力值低于 1(负相关)。② 较深色的瓷砖表示较强的相关性,无色瓷砖表示升力值约为 1($F_{(i)}$ 和 $R_{(j)}$ 之间的周相关性)。

3　结　语

本文提出了一种称为结构化关联映射图的新型可视化方法,精心设计来表示大型交易数据中项集之间的复杂关系。其与经典簇热图相似,因为矩阵与两个树形图组合。然而,结构化关联映射的树形图以更复杂的方式构建,以避免对多对多关联规则的误解。优化了之前工作中引入的结构化关联映射的抽象概念,并开发了构建优化其详细概念和增强的过程。

参　考　文　献

[1] Boulware L E, Barnes G J, Wilson R F, et al. Value of the periodic health evaluation[J]. Evidence Report/technology Assessment, 2006(136):1.
[2] Kweon S, Kim Y, Jang M J, et al. Data resource profile: the Korea National Health and Nutrition Examination Survey (KNHANES)[J]. International Journal of Epidemiology, 2014, 43(1):69-77.
[3] 边根庆,王月. 一种基于矩阵和权重改进的 Apriori 算法[J]. 微电子学与计算机,2017,34(1):136-140.
[4] 李春青. 基于关联规则算法的数据挖掘研究[J]. 软件导刊,2017,16(2):147-149.
[5] 肖晗,黄诚. 基于量化关联规则的敏感性分析[J]. 计算机应用,2017,37(S1):1-6.
[6] 刘晓蔚. 基于等价类规则树的高效关联规则挖掘算法[J]. 计算机应用与软件,2015,32(1):313-319.
[7] Yang L. Pruning and visualizing generalized association rules in parallel coordinates[J]. IEEE Transactions on Knowledge & Data Engineering, 2005, 17(1):60-70.
[8] Ferreira d O M C, Levkowitz H. From visual data exploration to visual data mining: a survey[J]. Visualization & Computer Graphics IEEE Transactions on, 2003, 9(3):378-394.
[9] Sekhavat Y A, Hoeber O. Visualizing Association Rules Using Linked Matrix, Graph, and Detail Views[J]. International Journal of Intelligence Science, 2013, 3(1):34-49.
[10] Hahsler M, Buchta C, Gruen B, et al. Mining Association Rules and Frequent Itemsets[EB/OL]. 2017. https://cran. r-project. org/web/packages/arules/arules. pdf.
[11] Tan P N, Steinbach M, Kumar V. Introduction to Data Mining, (First Edition)[M]. Addison-Wesley Longman Publishing Co. Inc. 2005.