

大数据环境下相容数据集的关联规则数据挖掘

张春生

(内蒙古民族大学 计算机科学与技术学院, 内蒙古 通辽 028043)

摘 要: 在对不可连接数据集充分分析的基础上, 给出基于不可连接数据集的相容数据集和不相容数据集 2 个定义, 给出相容数据集的一些基本理论, 在这些理论基础上给出一个基于相容数据集的关联规则挖掘方案, 实现每个相容数据集挖掘的关联规则直接合并, 生成整个相容数据集的关联规则, 实现普通数据挖掘算法无法实现的关联规则挖掘. 方案扩展了关联规则算法的应用领域, 提高了数据挖掘效率, 在一定程度上也实现了隐私保护.

关键词: 大数据; 相容数据集; 不相容数据集; 关联规则; 数据挖掘

中图分类号: TP311

文献标识码: A

文章编号: 1000—7180(2016)08—0034—06

Data Mining Algorithm of Association Rules Among Compatible Datasets in Big Data Environment

ZHANG Chun-sheng

(College of Computer Science and Technology, Inner Mongolia University for Nationalities, Tongliao 028043, China)

Abstract: On the basis of the data set which can not be connecteel, the definitions of the compatible data set and incompatible data set 2 based on the data sets not to be connected are given, some basic theories about compatible data set are presented, an association rule mining programme based on compatible data set is proposed. It is realized that association rules of compatible data set mining merge directly, so that association rules of the whole compatible data set are generated and the algorithm can realize an association rule mining which cannot be realized by the common data mining. The algorithm expands the application field of association rule algorithm and improves the efficiency of data mining, meanwhile, it realized the privacy protection in a certain extent.

Key words: big data; compatible data set; incompatible data set; association rule; data mining

1 引言

大数据是指数据量大, 产生的速度快, 数据的来源复杂, 有结构化、半结构化和无结构的数据, 同时由于隐私保护问题也增加了原始数据的复杂性^[1], 如何在大量繁杂的数据中去挖掘可用信息具有挑战性.

针对数据源的复杂性和数据的可用性, 研究者给出了很多有价值的研究成果, 李建中针对于大数据的可用性及对策给出了比较完整的论述^[2], 余伟等给出了基于 Web 大数据环境下的不一致跨源数据发现方法^[3], 文献^[4]给出了在缺失元组的情

况下数据分析的完整性方法, 这些研究都要对不一致性的大数据适当的预处理, 都要付出时间代价^[5].

数据挖掘是大数据分析的主要手段之一, 包括分类、估值、预言、关联规则、聚集、描述和可视化等, 可见关联规则仍然是数据挖掘最简单有效的方法之一^[6-7].

目前有关大数据的应用方法主要是采用 hadoop 平台, 数据分析采用分布式方法, 研究的热点基本上都是在减少数据的复杂性和提高数据挖掘效率两个方面^[8].

综上所述, 关联规则分析仍然是大数据时代的最简单有效数据挖掘方法之一, 一般都是采用统一

结构的数据,在分布式平台上进行的.但由于数据来源的复杂性,数据的结构可能不一致,需要大量的数据预处理时间,特别是由于隐私保护的影响,有时数据的整合几乎不可能.同时由于分布式计算需要各个节点间不定期进行数据的交换,影响了数据挖掘的速度.

文献[9]关于相容数据集已经给出了一些初步成果,但定义和一些关键的概念不够严谨,给出的 3 个定理也不具完整的系统性,本文在此基础上,给出更加严谨的相容数据集、不相容数据集和相关数据集的概念,进一步给出一套完整的相容数据集关联规则演绎策略,实现相容数据集关联规则演绎.

2 关联规则演绎理论

本文把数据集定义为 2 个大类,4 个小类.

相关数据集	单一数据集
	可连接数据集
不相关数据集	相容数据集
	不相容数据集

定义 1 属性一一对应,且对应属性具有相同域的属性组称为相容属性组.

定义 2 可通过关键字实现连接的多数据集或单一一个数据集称为相关数据集.

定义 3 不能通过关键字实现连接的数据集称为不相关数据集.

定义 4 存在一个及多个相容的属性组的两个不相关数据集称为相容数据集.

定义 5 不存在相容属性组的两个不相关数据集称为不相容数据集.

定义 6 关联规则 r_1 的最小支持度 s_1 ,最小置信度 c_1 ,关联规则 r_2 的最小支持度 s_2 和最小置信度 c_2 ,设 r_1 和 r_2 来源于两个相容数据集,若 $r_1 \theta r_2 \rightarrow r_3$,其中 θ 为演绎计算,则称 r_3 为演绎关联规则, r_3 支持度和置信度为 r_1 和 r_2 的最小值,即 $s = \min\{s_1, s_2\}$ 、 $c = \min\{c_1, c_2\}$.

文献[9]给出了定理 3、定理 4、定理 5 三个规则演绎定理,但不系统、不完整、不规范,本文在此基础上规范这三个定理,同时给出定理 4,并证明其正确性,从规范的角度给出整个相容数据集下的关联规则演绎方案.

定理 1 (右合并规则) 若关联规则 $r_1:A \rightarrow B$, $r_2:A \rightarrow C$ 是两个相容数据集上的关联规则, $r_1:A$ 和 $r_2:A$ 是相容属性组,则 r_1 和 r_2 右端可合并,即演绎关联规则 $r:A \rightarrow BC$ 成立.

定理 2 (左合并规则) 若关联规则 $r_1:B \rightarrow A$, $r_2:C \rightarrow A$ 是两个相容数据集上的关联规则, $r_1:A$ 和 $r_2:A$ 是相容属性组,则 r_1 和 r_2 左端可合并,即演绎关联规则 $r:BC \rightarrow A$ 成立.

定理 3 (传递规则) 若关联规则 $r_1:A \rightarrow B$, $r_2:B \rightarrow C$ 是两个相容数据集上的关联规则, $r_1:B$ 和 $r_2:B$ 是相容属性组,传递规则 $r:A \rightarrow C$ 成立条件为:

(1) 当 $s_1 > s_2$ 时,规则 $r:A \rightarrow BC$ 和 $r:A \rightarrow C$ 成立,当且仅当支持度 $\geq s$,置信度 $\geq c_1 \times c_2$.

(2) 当 $s_1 < s_2$ 时,规则 $r:A \rightarrow BC$ 和 $r:A \rightarrow C$ 成立.

(3) 当 $s_1 = s_2$ 时,规则 $r:A \rightarrow BC$ 和 $r:A \rightarrow C$ 成立.

定理 4 (一般合并规则) 若关联规则 $r_1:A \rightarrow BC$, $r_2:D \rightarrow BE$ 是两个相容数据集上的关联规则, $r_1:B$ 和 $r_2:B$ 是相容属性组, r_1 和 r_2 存在演绎规则: $r:A \rightarrow BCDE$ 和 $r:D \rightarrow ABCE$ 条件为:

(1) 当 $s_1 > s_2$ 时,规则 $r:D \rightarrow ABCE$ 成立.

(2) 当 $s_1 < s_2$ 时,规则 $r:A \rightarrow BCDE$ 成立.

(3) 当 $s_1 = s_2$ 时,规则 $r:D \rightarrow ABCE$ 和 $r:A \rightarrow BCDE$ 都成立.

证明:

(1) 支持度

由于 L_1 和 L_2 长度可能不同,由推论 1,可通过对某 1 个数据集进行比例缩放 $L_1 = L_2 = L$, L 为两个数据集的事务数. 构建 $s_1 > s_2$, $s_1 < s_2$, $s_1 = s_2$ 三种情形,分别如图 1、图 2、图 3 所示.

R_1			R_2		
A	B	C	D	B	E
...	...	...	...	...	...
...	...	...	...	...	E
A	B	C	D	B	
...	...	...	...	...	

图 1 $s_1 > s_2$

R_1			R_2		
A	B	C	D	B	E
...	...	...	...	...	...
...	...	C	...	...	...
A	B		D	B	E
...	...	...	...	...	...

图 2 $s_1 < s_2$

① 当 $s_1 > s_2$ 时,由图 1 可知,项集 $\{ABCDE\}$ 出现的次数与项集 $\{DBE\}$ 出现的次数相同,所以:

$$\text{supprot}(D \rightarrow ABCE) = \text{supprot}(A \rightarrow BCDE)$$

R_1			R_2		
A	B	C	D	B	E
...	...	...	...	...	...
...	...	...	...	...	E
A	B	C	D	B	
...	...	...	...	...	

图 3 $s_1 = s_2$

$$= \frac{\text{supprot}(ABCDE)}{L}$$

由于项集 $\{ABCDE\}$ 出现的次数与项集 $\{DBE\}$ 出现的次数相同,所以:

$$\begin{aligned} \text{supprot}(D \rightarrow ABCE) &= \text{supprot}(A \rightarrow BCDE) \\ &= \frac{\text{supprot}(DBE)}{L} \end{aligned}$$

$$\therefore \frac{\text{supprot}(DBE)}{L} = \text{supprot}(D \rightarrow BE) = s_2$$

$$\begin{aligned} \therefore \text{supprot}(D \rightarrow ABCE) \\ &= \text{supprot}(A \rightarrow BCDE) = s_2 \end{aligned}$$

$$\therefore s = \min\{s_1, s_2\}$$

$$\begin{aligned} \therefore \text{supprot}(D \rightarrow ABCE) &= \text{supprot} \\ (A \rightarrow BCDE) &\geq s \end{aligned}$$

② 当 $s_1 < s_2$ 时,由图 2 可知,项集 $\{ABCDE\}$ 出现的次数与项集 $\{ABC\}$ 出现的次数相同,所以:

$$\begin{aligned} \text{supprot}(D \rightarrow ABCE) &= \text{supprot}(A \rightarrow BCDE) \\ &= \frac{\text{supprot}(ABCDE)}{L} \end{aligned}$$

由于项集 $\{ABCDE\}$ 出现的次数与项集 $\{ABC\}$ 出现的次数相同,所以:

$$\begin{aligned} \text{supprot}(D \rightarrow ABCE) &= \text{supprot}(A \rightarrow BCDE) \\ &= \frac{\text{supprot}(ABC)}{L} \end{aligned}$$

$$\therefore \frac{\text{supprot}(ABC)}{L} = \text{supprot}(A \rightarrow BC) = s_1$$

$$\begin{aligned} \therefore \text{supprot}(D \rightarrow ABCE) \\ &= \text{supprot}(A \rightarrow BCDE) = s_1 \end{aligned}$$

$$\therefore s = \min\{s_1, s_2\}$$

$$\therefore \text{supprot}(D \rightarrow ABCE) = \text{supprot}(A \rightarrow BCDE) \geq s$$

③ 当 $s_1 = s_2$ 时,由图 3 可知,项集 $\{ABCDE\}$ 出现的次数与项集 $\{ABC\}$ 、 $\{DBE\}$ 出现的次数相同,所以:

$$\begin{aligned} \text{supprot}(D \rightarrow ABCE) &= \text{supprot}(A \rightarrow BCDE) \\ &= \frac{\text{supprot}(ABCDE)}{L} \end{aligned}$$

由于项集 $\{ABCDE\}$ 出现的次数与项集 $\{ABC\}$ 、 $\{DBE\}$ 出现的次数相同,所以:

$$\begin{aligned} \text{supprot}(D \rightarrow ABCE) &= \text{supprot}(A \rightarrow BCDE) \\ &= \frac{\text{supprot}(ABC)}{L} \\ &= \frac{\text{supprot}(DBE)}{L} \end{aligned}$$

$$\therefore \frac{\text{supprot}(ABC)}{L} = \text{supprot}(A \rightarrow BC) = s_1$$

$$\therefore \frac{\text{supprot}(DBE)}{L} = \text{supprot}(D \rightarrow BE) = s_2$$

$$\begin{aligned} \therefore \text{supprot}(D \rightarrow ABCE) &= \text{supprot}(A \rightarrow BCDE) \\ &= s_1 = s_2 \end{aligned}$$

$$\therefore s = \min\{s_1, s_2\}$$

$$\therefore \text{supprot}(D \rightarrow ABCE) = \text{supprot}(A \rightarrow BCDE) \geq s$$

(2) 置信度

① 当 $s_1 > s_2$ 时,由图 1 可知,项集 $\{ABCDE\}$ 出现的次数与项集 $\{DBE\}$ 出现的次数相同,所以:

$$\begin{aligned} \text{con}(D \rightarrow ABCE) &= \frac{\text{supprot}(ABCDE)}{\text{support}(D)} \\ &= \frac{\text{supprot}(DBE)}{\text{support}(D)} \end{aligned}$$

$$\frac{\text{supprot}(DBE)}{\text{support}(D)} = \text{con}(D \rightarrow BE) = c_2$$

$$\therefore c = \min\{c_1, c_2\}$$

$$\therefore \text{con}(D \rightarrow ABCE) = \text{con}(D \rightarrow BE) \geq c$$

② 当 $s_1 < s_2$ 时,由图 2 可知,项集 $\{ABCDE\}$ 出现的次数与项集 $\{ABC\}$ 出现的次数相同,所以:

$$\begin{aligned} \text{con}(A \rightarrow BCDE) &= \frac{\text{supprot}(ABCDE)}{\text{support}(A)} \\ &= \frac{\text{supprot}(ABC)}{\text{support}(A)} \end{aligned}$$

$$\frac{\text{supprot}(ABC)}{\text{support}(A)} = \text{con}(A \rightarrow BC) = c_1$$

$$\therefore c = \min\{c_1, c_2\}$$

$$\therefore \text{con}(A \rightarrow BCDE) = \text{con}(A \rightarrow BC) \geq c$$

③ 当 $s_1 = s_2$ 时,由图 3 可知,项集 $\{ABCDE\}$ 出现的次数与项集 $\{ABC\}$ 、 $\{DBE\}$ 出现的次数相同,所以:

$$\begin{aligned} \text{con}(D \rightarrow ABCE) &= \frac{\text{supprot}(ABCDE)}{\text{support}(D)} \\ &= \frac{\text{supprot}(DBE)}{\text{support}(D)} \end{aligned}$$

$$\frac{\text{supprot}(DBE)}{\text{support}(D)} = \text{con}(D \rightarrow BE) = c_2$$

$$\therefore c = \min\{c_1, c_2\}$$

$$\therefore \text{con}(D \rightarrow ABCE) = \text{con}(D \rightarrow BE) \geq c$$

$$\text{con}(A \rightarrow BCDE) = \frac{\text{supprot}(ABCDE)}{\text{support}(A)}$$

$$= \frac{\text{supprot}(ABC)}{\text{supprot}(A)}$$

$$\frac{\text{supprot}(ABC)}{\text{supprot}(A)} = \text{con}(A \rightarrow BC) = c_1$$

$$\because c = \min\{c_1, c_2\}$$

$$\therefore \text{con}(A \rightarrow BCDE) = \text{con}(A \rightarrow BC) \geq c$$

(3) 讨论

① 当 $s_1 > s_2$ 时

$$\text{supprot}(D \rightarrow ABCE) = \text{supprot}(A \rightarrow BCDE) \geq s$$

$$\text{con}(D \rightarrow ABCE) = \text{con}(D \rightarrow BE) \geq c$$

所以,规则 $r: D \rightarrow ABCE$ 成立.

② 当 $s_1 < s_2$ 时

$$\text{supprot}(D \rightarrow ABCE) = \text{supprot}(A \rightarrow BCDE) \geq s$$

$$\text{con}(A \rightarrow BCDE) = \text{con}(A \rightarrow BC) \geq c$$

所以,规则 $r: A \rightarrow BCDE$ 成立.

③ 当 $s_1 = s_2$ 时

$$\text{supprot}(D \rightarrow ABCE) = \text{supprot}(A \rightarrow BCDE) \geq s$$

$$\text{con}(A \rightarrow BCDE) = \text{con}(A \rightarrow BC) \geq c$$

所以,规则 $r: D \rightarrow ABCE$ 和 $r: A \rightarrow BCDE$ 都成立.

定理 4 得证.

3 时间复杂度分析

由于相容数据集是不可连接的,所以无法对比分析时间复杂度,因此本文只有与可连接数据集进行对比分析.

设数据集 D_1 和数据集 D_2 为两个相容数据集,假设是可连接的,由于是量级比较,不妨假设 D_1 和数据集 D_2 中事务数均为 $|T|$,事务中的属性数目均为 n , $|L_k|$ 是频繁 k 项集数量, $|C_k|$ 是候选 k 项集数量,假设最大频繁项集能达到 n ,下面对本算法时间复杂度进行分析.

由文献[10]得到 Apriori 算法时间复杂度为:

$$o(n |T| + \sum_{2 \leq k \leq n} (|L_k| \times |L_k| + |T| \times |C_k| + |C_k|)) \quad (1)$$

(1) D_1 与 D_2 连接数据集挖掘时间复杂度

此时数据集中事务数为 $|T|$,属性数为 $2n$,时间复杂度为:

$$o(2n |T| + \sum_{2 \leq k \leq 2n} (|L_k| \times |L_k| + |T| \times |C_k| + |C_k|)) \quad (2)$$

(2) D_1 与 D_2 独立挖掘复杂度

此时数据集中事务数为 $|T|$,属性数为 n ,时间复杂度为:

$$o(2n |T| + 2 \sum_{2 \leq k \leq n} (|L_k| \times |L_k| + |T| \times |C_k| + |C_k|)) \quad (3)$$

(3) 两种算法的复杂度差

$$o(\sum_{2 \leq k \leq 2n} (|L_k| \times |L_k| + |T| \times |C_k| + |C_k|) - 2 \sum_{2 \leq k \leq n} (|L_k| \times |L_k| + |T| \times |C_k| + |C_k|)) \quad (4)$$

假设最大频繁项集能达到 n ,此时两种算法的复杂度差为:

$$o(\sum_{n+1 \leq k \leq 2n} (|L_k| \times |L_k| + |T| \times |C_k| + |C_k|) - \sum_{2 \leq k \leq n} (|L_k| \times |L_k| + |T| \times |C_k| + |C_k|)) \quad (5)$$

(4) 复杂度讨论

由于本算法是在单独数据集上挖掘出规则后直接进行规则合并,中间不进行数据交换,所以优于分布式挖掘算法;从式(5)可以看出,本算法明显优于连接数据集挖掘算法.

4 关联规则演绎方案

4 个规则除可单独进行规则演绎外,还可灵活运用 4 个规则可实现更复杂的混合运算.

例 1 设有规则 $r_1: A \rightarrow B$,支持度为 s_1 ,置信度为 c_1 ; $r_2: A \rightarrow C$,支持度为 s_2 ,置信度为 c_2 ; $r_3: D \rightarrow BE$,支持度为 s_3 ,置信度为 c_3 ,其中 $s_1 > s_2, s_2 > s_3$, $r_1: A$ 与 $r_2: A$ 为相容属性组, $r_1: B$ 与 $r_3: B$ 为相容属性组,求 r_1, r_2, r_3 的演绎关联规则.

解:

由定理 1 即右合并规则可知 r_1 和 r_2 可演绎出规则 $r_4: A \rightarrow BC$,支持度 $s_4 = \min\{s_1, s_2\}$,由于 $s_1 > s_2$,所以 $s_4 = s_2$.

由于 $s_2 > s_3, s_4 = s_2$,所以 $s_4 > s_3$,此时考虑 $r_4: A \rightarrow BC$ 和 $r_3: D \rightarrow BE$,由定理 4 即一般合并规则可知 $r_5: D \rightarrow ABCE$ 成立.

得到 2 个演绎规则 r_4, r_5 .

例 2 设有规则 $r_1: A \rightarrow B$,支持度为 s_1 ,置信度为 c_1 ; $r_2: A \rightarrow C$,支持度为 s_2 ,置信度为 c_2 ; $r_3: D \rightarrow BE$,支持度为 s_3 ,置信度为 c_3 ,其中 $s_1 > s_2, s_2 < s_3, r_1: A$ 与 $r_2: A$ 为相容属性组, $r_1: B$ 与 $r_3: B$ 为相容属性组,求 r_1, r_2, r_3 的演绎关联规则.

解:

由定理 1 即右合并规则可知 r_1 和 r_2 可演绎出规则 $r_4: A \rightarrow BC$,支持度 $s_4 = \min\{s_1, s_2\}$,由于 $s_1 > s_2$,所以 $s_4 = s_2$.

由于 $s_2 < s_3, s_4 = s_2$,所以 $s_4 < s_3$,此时考虑 $r_4:$

$A \rightarrow BC$ 和 $r_3: D \rightarrow BE$, 由定理 4 即一般合并规则可知规则 $r_5: A \rightarrow BCDE$ 成立.

得到 2 个演绎规则 r_4, r_5 .

例 3 设有规则 $r_1: A \rightarrow B$, 支持度为 s_1 , 置信度为 c_1 ; $r_2: A \rightarrow C$, 支持度为 s_2 , 置信度为 c_2 ; $r_3: B \rightarrow D$, 支持度为 s_3 , 置信度为 c_3 ; $r_4: C \rightarrow D$, 支持度为 s_4 , 置信度为 c_4 , 其中 $s_1 > s_2, s_3 > s_4, s_2 > s_4, r_1: A$ 与 $r_2: A$ 为相容属性组, $r_1: B$ 与 $r_3: B$ 为相容属性组, $r_2: C$ 与 $r_4: C$ 为相容属性组, $r_3: D$ 与 $r_4: D$ 为相容属性组, 求 r_1, r_2, r_3, r_4 的演绎关联规则.

解:

由定理 1 即右合并规则可知 r_1 和 r_2 可演绎出规则 $r_5: A \rightarrow BC$, 支持度 $s_5 = \min\{s_1, s_2\}$, 由于 $s_1 > s_2$, 所以 $s_5 = s_2$.

由定理 2 即左合并规则可知 r_3 和 r_4 可演绎出规则 $r_6: BC \rightarrow D$, 支持度 $s_6 = \min\{s_3, s_4\}$, 由于 $s_3 > s_4$, 所以 $s_6 = s_4$.

由于 $r_1: B$ 与 $r_3: B$ 为相容属性组, $r_2: C$ 与 $r_4: C$ 为相容属性组, 所以属性组 $r_5: \{BC\}$ 与 $r_6: \{BC\}$ 也相容.

由于 $s_5 = s_2, s_6 = s_4, s_2 > s_4$, 所以 $s_5 > s_6$, 由定理 3 即传递规则可知, r_5 和 r_6 可产生规则 $r_7: A \rightarrow BCD$ 和 $r_8: A \rightarrow D$.

得到 4 个演绎规则 r_5, r_6, r_7, r_8 .

5 实验分析

Movieclick 数据库是微软公司提供的用于数据挖掘的测试数据集, 来源于某公司员工的基本情况以及对电影的爱好和摄取方式的调查问卷. 数据库有一个基本表 customers 和 7 个卫星表构成星型结构, 基本信息大部分在基本表 customers 中.

Customers(customerID, age, education level, gender, home ownership, internet connection, marital status, movie selector, num bathrooms, num bedrooms, num cars, num children, num TVs, FPV freq, prerec buying freq, prerec format, prerec renting freq, theater freq, TV movie freq, TV signal). 其中含有 3 187 条记录.

为实现测试目的, 本文针对于重点属性, 在 Customers 表中进行全局数据挖掘, 得到感兴趣的全局关联规则, 同时将 Customers 表进行水平分割, 构建两个相容数据集(来源于 2 组不同的员工, 不能进行数据连接), 分别对这两个表进行挖掘, 得到感兴趣的局部关联规则, 最后通过局部关联规则合并,

得到在全局表中挖掘的关联规则相同的规则, 证明相容数据集关联规则的合并与全局关联规则是相互吻合的, 表明本文理论的正确性.

(1) 实验环境

硬件环境: 联想 PC, CPU Pentium(R), G645 @2.90 GHz

软件环境: windows 7, Visual Studio 2010, MS SQL Server 2008

(2) Customers 表的奇异点处理

Customers 中有 3 187 条记录, 实验重点探讨员工的学历与住房情况及子女数量的关系, 而数据库中有 57 条记录的 customerID、education level、home ownership、num children 4 个属性数据不全, 按奇异点处理方法, 删除这 57 条记录, 保留 3 130 条记录.

(3) 构建全局数据集和局部数据集

全局数据集: Customers(customerID, education level, home ownership, num children)

记录号: 1~3 130, 记录数: 3 130

局部数据集: Customers_1(customerID, education level, home ownership)

记录号: 1~1 565, 记录数: 1 565, 去掉属性 num children

Customers_2(customerID, education level, num children)

记录号: 1 566~3 130, 记录数: 1 565, 去掉属性 home ownership

(4) 数据预处理

由于学历的属性值有 9 个, num children 的属性值有 6 个, 过于分散, 不利于规则的产生, 所以要进行必要的数据预处理.

education level 的预处理:

(doctorate, master's degree, post-doc) 替换为“高”

(associate's degree, bachelor's degree, some college) 替换为“中”

(grade school, high school, trade school) 替换为“低”

home ownership 的预处理:

rent 替换为“租用”; own 替换为“自有”.

num children 的预处理:

{0, 1, 2} 替换为“少”; {3, 4} 替换为“中”; {5, 6} 替换为“多”.

(5) 挖掘结果

Customers:存在典型关联规则

$R: (\text{education level} = \text{“中”}) \rightarrow (\text{home ownership} = \text{“自有”}) (\text{num children} = \text{“少”})$

支持度 $s: 1\ 136/3\ 130 = 36.3\%$, 置信度 $c: 1\ 136/2\ 058 = 55.2\%$

Customers_1:存在典型关联规则

$R_1: (\text{education level} = \text{“中”}) \rightarrow (\text{home ownership} = \text{“自有”})$

支持度 $s_1: 652/1\ 565 = 41.6\%$, 置信度 $c_1: 652/991 = 65.8\%$

Customers_2:存在典型关联规则

$R_2: (\text{education level} = \text{“中”}) \rightarrow (\text{num children} = \text{“少”})$

支持度 $s_2: 946/1\ 565 = 60.4\%$, 置信度 $c_2: 946/1\ 067 = 88.7\%$

(6)规则演绎

按定理 1(右合并规则), R_1 与 R_2 合并, 得到新规则:

$R': (\text{education level} = \text{“中”}) \rightarrow (\text{home ownership} = \text{“自有”}) (\text{num children} = \text{“少”})$

支持度 $s' = \min\{41.6\%, 60.4\%\} = 41.6\%$, 置信度 $c' = \min\{65.8\%, 88.7\%\} = 65.8\%$

(7)结论

由于 R' 的支持度 $s' > s$, 置信度 $c' > c$, 所以合并规则 R' 成立, 并且与规则 R 相同。

本例通过对两个相容数据集的挖掘, 并对得到的关联规则进行合并, 得到的新关联规则与全局挖掘结果相吻合, 证明了本文的正确性。

6 结束语

提出一种大数据环境下相容数据集关联规则演绎方案. 该方案可单独在异构的相容数据集上进行数据挖掘, 不需要数据整合, 节省了数据预处理时间, 单独挖掘不需要挖掘过程中的数据交换, 提高了数据挖掘效率. 由于可不进行数据连接, 克服了数据隐私保护带来的限制, 同时由于可在仅有部分相容属性组的异构数据集上单独进行数据挖掘, 得到的规则直接进行演绎, 可实现普通数据挖掘算法无法

挖掘到的更深层的隐蔽关联规则。

在文献[9]基础上, 给出了更加严谨的定义和定理, 同时也给出了更一般的演绎规则, 文章最后通过案例形式给出了关联规则演绎方法, 通过对微软公司 Movieclick 数据库的实验更进一步证明了本方案的可行性。

参考文献:

[1] 刘雅辉, 张铁赢, 靳小龙, 等. 大数据时代的个人隐私保护[J]. 计算机研究与发展, 2015, 52(1): 221-228.

[2] 李建中, 刘显敏. 大数据的一个重要方面: 数据可用性[J]. 计算机研究与发展, 2013, 50(6): 1147-1162.

[3] 余伟, 李石君, 杨莎. Web 大数据环境下的不一致跨源数据发现[J]. 计算机研究与发展, 2015, 52(2): 295-308.

[4] Fan Wenfei, Geerts F. Relative information completeness[C]// Proc of the 28th ACM SIGMOD-SIGACT-SIGART Symp on Principles of Database Systems. New York: ACM, 2009: 97 -106

[5] 朱付保, 白庆春, 汤萌萌, 等. 基于改进 Apriori 算法的铁路轨道质量分析与评价[J]. 微电子学与计算机, 2015, 32(10): 159-162.

[6] 周丽娟, 王翔. 云环境下关联规则算法的研究[J]. 计算机工程与设计, 2014, 35(2): 499-503.

[7] Gupta R, Gupta H, Mohania M K. Cloud computing and big data analytics: What is new from databases perspective? //Proc of the 1st Int Conf of Big Data Analytics. Berlin, Springer, 2012: 42-61.

[8] 李敏, 倪少权, 邱小平. 物联网环境下基于上下文的 Hadoop 大数据处理系统模型[J]. 计算机应用, 2015, 35(5): 1267-1272.

[9] 张春生, 庄丽艳. 基于 Apriori 的相容数据集间关联规则演绎方法[J]. 计算机应用, 2013, 33(10): 2796-2800.

[10] 翁敬农. 数据挖掘教程[M]. 北京: 清华大学出版社, 2003.

作者简介:

张春生 男, (1965-), 硕士, 教授. 研究方向为数据库、数据挖掘、软件理论. E-mail: zhangcs_817@163.com.