

可拓数据挖掘方法在解决矛盾问题中的研究^①

An Extension Data Mining Method in Resolution of Contradictions

张海涛 董春游 (黑龙江科技学院 计算机与信息工程学院 黑龙江 哈尔滨 150027)

摘要: 阐述了可拓数据挖掘概念和矛盾问题的形式化模型,研究了可拓数据挖掘方法过程及解决矛盾问题的方法,并通过实例进行了说明,得出应用可拓数据挖掘解决矛盾问题的策略可以为技术实现策略生成探索出一条可行之路。

关键词: 可拓 关联函数 可拓模型 可拓数据挖掘 矛盾问题

1 可拓数据挖掘概念

可拓学是基于矛盾问题而提出,它以矛盾问题研究对象,建立了把问题进行形式化描述的模型,并利用事物的可拓性和可拓变换,建立了解决矛盾问题的可拓方法和可拓工程方法,同时也为数据挖掘提供了新的研究方法和工具^[1]。

可拓数据挖掘是指为了处理矛盾和不相容问题在数据挖掘中引入可拓变换方法,并在数据挖掘获得静态知识的基础上,通过可拓变换,获取变化知识,即包含可拓变换的规则知识。它与传统数据挖掘的差异之一是:传统数据挖掘是知识的发现,而可拓数据挖掘不但可以挖掘知识,而且可以挖掘相关规则和可拓变换,从而挖掘变化的知识,因此它是策略生成的重要工具^[2]。

2 矛盾问题的形式化模型

可拓学把矛盾问题分为不相容问题和对立问题。

1) 不相容问题

给定目标 G 和条件 L , 称它们构成问题 P , 记作 $P=G*L$, 其中, G, L 分别为基元、基元的运算式或它们的组合, 简称为目标元和条件元, 在问题 $P=G*L$ 中, 当条件 L 能使目标 G 实现时, 称问题为相容问题, 记作 $G \downarrow L$; 否则, 称问题为不相容问题, 记作 $G \uparrow L$ 。

不相容问题属于主客观矛盾问题, 即客观条件下不能满足主观愿望的要求。

2) 对立问题

设问题 P 由两个目标 G_1, G_2 和条件 L 构成, 即 $P=(G_1 \wedge G_2)*L$ 在条件 L 下, 两个目标能同时实现, 称问题 P 为共存问题, 记作 $(G_1 \wedge G_2) \downarrow L$; 反之记作 $(G_1 \wedge G_2) \uparrow L$ 。

即同一条件下两个目标不能同时实现。

3 可拓数据挖掘过程及解决矛盾问题的方法

3.1 可拓数据挖掘过程^[3]

对于两类规则:

$$A \rightarrow P \quad (1)$$

$$B \rightarrow N \quad (2)$$

一般情况 $A = \bigwedge a_i, B = \bigwedge b_i$ 。

若存在条件的可拓变换

$$T \text{ 条件}(B) = A \quad (3)$$

并存在结论的可拓变换 T 结论(它为 T 条件的传导变换),

$$T \text{ 结论}(N) = P \quad (4)$$

则可拓变换规则知识(变化知识)

$$T_B(B) = A \rightarrow T_N(N) = P \quad (5)$$

成立, 即

$$\text{if } T_B(B) = A \text{ then } T_N(N) = P \quad (6)$$

从可拓数据挖掘定理中, 可以概括可拓数据挖掘过程为:

Step 1 对分类问题利用数据挖掘方法获得分类规则, 即获得式(1)和式(2)的知识。

^① 收稿时间:2008-09-24

Step 2 确定规则的前提中存在的可拓变换以及结论中存在的可拓变换,即找出满足式(3)和式(4)的可拓变换。

Step 3 获得可拓知识式(5)。

3.2 可拓数据挖掘方法生成解决矛盾问题策略的一般步骤

解决矛盾问题必须通过可拓变换,即利用变化的信息才能解决矛盾问题。^[4]其工具是关联函数:关联函数公式 $k(x)=\rho(x,X_0)/D(x,X_0,X)$, 其中 $D(x,X_0,X)$ 是点 x 关于区间套 X_0 和 X 的位值, $\rho(x,X_0)$ 代表点 x 与区间 $X_0=<a,b>$ 之距, $X_0=<a,b>$ 。关联函数本身属于知识。若 $k_0(P)<0$, 即问题 P 为不相容问题, 如存在变换 T 使 $T_k k(Tl_0l_0)=k' (Tl_0l_0)=k' (l_0')>0$, 则称问题 P 得到了解决。

对矛盾问题 $P=G*L$, 可拓数据挖掘主要是挖掘它的解变换.应用可拓数据挖掘方法生成解决矛盾问题的策略,即解变换的一般步骤是:

1)建立问题的可拓模型和确定相关规则的类型

首先要界定问题并建立问题在于可拓模型 $P=G*L$, 然后确定核问题, 记 $P_0=g_0*l_0$ 为问题 P 的核问题, 其中 $g_0=(Z_0,c_0s,X_0), l_0=(Z_0,c_0t,c_0t(Z_0))$. 然后运用相关性对核问题进行相关分析, 建立问题的相关规则。

2)建立基础数据表 B

在关系数据库中, 与相关规则相关的数据集可以通过选择、投影、连接等操作来实现。

假定与规则相关的数据表为 $b_1,b_2,\dots,b_k$, 可通过 SQL 建立基础数据表 B 。然后通过分类, 同时增加支持度、可信度与关联函数而获取基础规则库 DW 。

3)建立强相关规则库 KW , 在基础规则库 DW 的基础上, 将支持度 $\geq \text{minsupport}$ 与可信度 $\geq \text{minconfidence}$ 的规则存入强相关规则库 KW 。

4)建立矛盾问题库, 将关联函数值 $k \leq 0$ 的问题存入问题库 QW , 并按关联函数值 k 升序排序。

5)策略生成与评价

由于强相关规则表示的是成功的经验与策略, 生成策略的基本方法之一是从强相关规则库中匹配合适的规则。

对于从矛盾问题库中选择出来的问题, 进行种种变换后, 可得到矛盾问题的解变换集, 即策略集, 这

些策略的优度不同, 必须对它们进行评价选优, 选出优度较高的若干策略作为待实施策略。通常使用的方法是优度评价法, 公式为:

$$c(T_j)=a_jk_j[M_j(T_j)](j=1,2,\dots,m).$$
 较佳策略为 T_0 :
$$C(T_0)=\max(C(T_j)|j=1,2,\dots,m)$$

4 可拓数据挖掘方法应用实例

某学生 A 想把学习成绩在现在的基础上提高 10%, 这显然是一个主客观矛盾问题, 现以此做为一个实例来证明可拓数据挖掘方法在解决该矛盾问题中的应用:

4.1 首先建立可拓模型

把一段时间内综合评定成绩增长不少于 10%作为目标, 建立可拓模型如下, 即

$$P=I*r=\begin{bmatrix} \text{提高成绩} & \text{支配对象} & \{\text{学生A}\} \\ & \text{比例} & \geq 10\% \end{bmatrix} * \begin{bmatrix} \{\text{学生A}\} & \text{最高成绩} & M \\ & \text{现在成绩} & N \end{bmatrix}$$

选择相应的关联函数: $k(x)=\frac{\rho_r(x,x_0,X_0)}{D(x,X_0,X)}$
(具体过程略)。

根据教师经验, 学生成绩状况主要与智力因素、学习方法、学习目标、学习动机有关。因此, 每个学生用多维物元来表示: 即

$$R_{ijk}=\begin{bmatrix} \text{学生} & \text{智力因素} & C_1 & V_{ijk1} \\ & \text{学习方法} & C_2 & V_{ijk2} \\ & \text{学习目标} & C_3 & V_{ijk3} \\ & \text{学习动机} & C_4 & V_{ijk4} \\ & \text{学习成绩} & C_5 & V_{ijk5} \end{bmatrix}=\begin{bmatrix} R_1 \\ R_2 \\ R_3 \\ R_4 \\ R_5 \end{bmatrix}$$

问题 P 的核问题为: $P_0=g_0*l_0=(\{\text{学生 A}\}, \text{目标成绩}, [1.1N,2N])*(\{\text{学生 A}\}, \text{现在成绩}, N)$ 。

4.2 相关分析

从上面的模型可以看出, 要提高学习成绩, 必须综合改善多个因素, 故可以表述为:

$$R_1 \wedge R_2 \wedge R_3 \wedge R_4 = > (I) R_5.$$
 本实例中, 取学生 A 的成绩可信度 $\text{confidence}=\text{考试得分}/\text{学科总分}$ 。

4.3 建立基础数据表 B

可通过 SQL 建立基础数据表,如下表:

4.4 建立强相关规则库(KW)

将支持度 $\geq \text{minsupport}$ 与支持度 $\geq \text{minconf} - \text{idence}$ 的规则存入强相关规则库。

4.5 建立矛盾问题库(QW)

将关联函数值 $k \leq 0$ 的规则对应的问题选为要分析的矛盾问题，并存入矛盾问题库，并按关联函数值 k 升序排序。

表 1 基础数据表(部分)

序号	条件 1 智力因素	条件 2 学习方法	条件 3 学习目标	条件 4 学习动机	结果 综合成绩	支持度	可信度	关联函数 k
1	120	优	明确	中	优	0.215	0.567	0.067
2	110	良	模糊	高	优	0.113	0.704	0.204
3	100	差	模糊	低	一般	0.025	0.333	-0.167

问题库的部分内容如下表 2:

表 2 实例问题数据表(部分)

序号	条件 1 智力因素	条件 2 学习方法	条件 3 学习目标	条件 4 学习动机	结果 综合成绩	支持度	可信度	关联函数 k
5	100	差	模糊	低	一般	0.025	0.333	-0.167
7	90	中	模糊	中	差	0.038	0.333	-0.167

4.6 策略生成与评价

4.6.1 实施基本变换.

对问题进行拓展推理，在初始态相关树中，即是对“叶”基元进行拓展推理.可以应用发散推理或可扩

推理.而共轭推理只有在初始态相关树中的“叶”基元为物元时才可应用.现从问题库中选择一个问题为例:

表 3 实例问题相关规则数据表(部分)

序号	条件 1 智力因素	条件 2 学习方法	条件 3 学习目标	条件 4 学习动机	结果 综合成绩	支持度	可信度	关联函数 k
22	100	差	模糊	低	差	0.025	0.333	-0.167

本例中选择序号为 22 的实例问题来分析，用相关规则表示：

{学生A },智力因素, 100
学习方法, 差
学习目标, 模糊
学习动机, 低

→(0.025,0.333, -0.167)

{学生 A}, 综合成绩, 差),

设= {学生A },智力因素, 100
学习方法, 差
学习目标, 模糊
学习动机, 低

则叶基元的集合={ $r_{13},r_{23},r_{36},r_{41}$ }.这里用“一特征多值”进行发散，即 $r_{23} - \{r_{21},r_{22},r_{24}\}$,相应地，可拓变换为 $T_{231}r_{23}=r_{21}$, $T_{232}r_{23}=r_{22}$, $T_{234}=r_{24}$.

令 $T_1 = (e, T_{232}, e, e)$, $T1 \in \{T|T:QW \rightarrow KW\}$,则变换 T_1 为矛盾问题的解变换。即：

{学生A },智力因素, 100
学习方法, 一般
学习目标, 模糊
学习动机, 低

→ (0.025,0.667,0.167)

{学生 A},成绩, 良),实施基本变换后，达到目标。

4.6.2 变换的运算

令 $T_2 = (e, T_{231}, e, T_{412})$, $T_2 \in \{T|T:QW \rightarrow KW\}$,则变换 T_2 为矛盾问题的解变换，也可以令 $T_3 =$

($e, T_{231}, T_{365}, T_{412}$), $T_3 \in \{T | T: QW \rightarrow KW\}$ 为矛盾问题的解变换。

下面把问题的三种解变换列表如下:

表 4 问题与变换

相 关 规 则 类 型	条 件 1 智 力 因 素	条 件 2 学 习 方 法	条 件 3 学 习 目 标	条 件 4 学 习 动 机	结 果 综 合 成 绩	支 持 度	可 信 度	关 联 函 数 k
问题	100	差	模 糊	低	差	0.025	0.333	-0.167
变换 T_1	100	一 般	模 糊	低	良	0.025	0.667	0.167
变换 T_2	100	好	模 糊	中	良	0.075	0.667	0.167
变换 T_3	100	好	明 确	中	优	0.075	0.833	0.333

本例中生成了解决矛盾问题的策略,即想改变智力因素不容易,通过优度评价法可得出变换 T_3 是最优先考虑的变换,也就是改善学习方法,其次是学习目标,最后才是学习动机。此三种变换都可以使关联函数 K 从小于 0 变为大于 0,使问题得到解决。

5 结束语

可拓数据挖掘方法是一种新的方法,其理论正不断发展,本文尝试用可拓数据挖掘方法应用于解决不相容问题的研究,并取得常规方法很难取得的结果,现在做的是比较初步的工作,将来本文的研究会应用于一种新的可拓策略生成系统之中。

参考文献

- 1 李立希,杨春燕.可拓策略生成系统.北京:科学出版社,2006:133-142.
- 2 杨春燕,蔡文.可拓工程.北京:科学出版社,2007:261.
- 3 陈文伟,黄金才.可拓知识与可拓数据挖掘.广西师范大学学报:自然科学版,2006,24(4):160.
- 4 陈文伟.挖掘变化知识的可拓数据挖掘研究.中国工程科学,2006,8(11):70-714.