不完备知识下的形式概念表示与计算*)

吴 强[1,2]

(绍兴文理学院计算机系 绍兴 312000)[1] (上海大学计算机学院 上海 200072)[2]

摘 要 从集合的角度来说,知识就是数据集合在某种关系下的划分。如果这个数据集的某些属性值是未知的或丢失了,那么知识就是不完备(incomplete)的。传统形式概念分析是源于完备数据集的(完备知识)。在不完备知识下的概念分析一般说来比完备知识更困难。本文提出了一个新的不完备知识下形式概念表示与计算的方法,这种方法是基于泛化粗糙集理论的,其目的是扩展形式概念分析研究的领域。文中研究了一个基于自反相似关系的粗糙集模型,讨论了基于这种模型的形式概念分析方法。一个实例表明了这种方法的可行性。

关键词 不完备知识,粗糙集,形式概念,概念表示与计算

Representation and Computing of Formal Concepts under Incomplete Knowledge

WU Qiang[1,2]

(Department of Computer Science, Shaoxing University, Zhejiang 312000)[1]

(School of Computer Engineering and Science, Shanghai University, Shanghai 200072)[2]

Abstract From the point of view of the set, knowledge is a partition of data set about some relations. It is incomplete if some attribute values are unknown or missing in the data set. The classical formal concept analysis derives from complete data set (complete knowledge). Concept analysis under incomplete knowledge is usually more difficult. This paper presents new approaches to represent and compute formal concepts under incomplete knowledge. It is based on generalized rough set theory, and aims to extend formal concept analysis to the condition of incomplete knowledge. A rough set model on reflexive similarity relation is discussed. The approaches of formal concept analysis, build on the model, are proposed. Finally an example is given to show the feasibility of the proposed method.

Keywords Incomplete knowledge, Rough set, Formal concept, Concept representation and computing

1 引言

概念作为知识的基本单元,很自然地成为人工智能的重要研究对象。在哲学中,概念被理解为由外延和内涵两个部分所组成的思想单元。基于概念的这一理解,Wille[1]提出了形式概念分析(FCA)的思想。形式概念分析是一个在信息科学中十分有用的数学方法,已被应用于知识获取、表示和组织等[15]。这种应用的前提是研究的数据是确定的、明晰的。但是在经济、工程、环境、社会科学和医学等许多复杂的问题中,由于人为因素或客观条件的限制等,得到的数据并不总是确定和清晰的。在这种情况下,往往需要解决的一个重要问题就是形式概念的表示与计算。

从集合的角度来说,知识就是数据集合在某种关系下的划分。如果这个数据集的某些属性值是未知的或缺失,那么知识就是不完备的。传统形式概念分析是源于完备知识的。但是即使在完备知识下,并不是每个概念都是可以表示的。被研究的问题如果不能够用形式概念描述,它所涉及的概念被称为不可定义概念。发现最优描述不可定义概念的过程称为概念近似。概念近似在文[5]中第一次提出,后又在文[9]中得到了进一步研究。概念近似是以粗糙集作为一个基本的近似模型,所依据的粗糙集理论[2]是基于完备信息系统的,被处理的数据是完备的。文[6]给出了一个不完备知识的属性确定(attribute exploration)算法,它是比较确切的针对不完备知识的形式概念分析的研究。这个算法可以帮助用户得到未知背景有效属性隐含的关于其知识的最大信息量。文[7,13]进行了进一步研究,但它并没有讨论形式概念的表示问题。

一般来说,信息系统是指信息和知识表达系统,其本质也是一个对象属性关系的二维表。不完备信息系统属性值确实的粗糙集方法在文[8]中首先提出,尽管在文[16]中也简单提到过。类似于完备信息系统用等价不可分辨关系描述,不完备信息系统使用相似特征关系来刻画[11]。对于完备信息系统来说,一旦不可分辨关系确定并且概念(一个实例集)给出,上下近似就是唯一的。对给定的特征关系和概念,不完备信息系统有3种可能定义上下近似的方法:单元集(singleton)、子集(subset)和概念近似(concept approximations)[11]。文[12,17~20]研究了单元集上下近似。类似的思想也在文[21~25]体现出来。文[8]进一步讨论了3种方法在数据挖掘中的应用。

尽管有众多的不完备知识的处理方法,但从形式概念分析的角度开展的研究并不多见。

本文论证了一个新的不完备知识的概念表示方法,它是基于粗糙近似的泛化定义的[4]。与传统定义不同,当相似关系是自反的但不一定是对称和传递的情况下,这一定义正确地刻画了正对象集(下近似)与正或不确定对象集(上近似)。

*)国家自然科学基金项目"面向本体的形式概念分析理论和算法"(项目编号:60275022)。吴 强 副教授,博士生,主要研究方向:知识发现与知识表示。

由于它降低了构成关系的条件，能有效地解决知识中的不确定和缺失值问题，进而完成形式概念在不完备知识条件下的表示与计算。这种方法对处理大背景、多不确定或缺失值的问题尤为有效。

本文以下部分是这样组织的：第 2 节中提出了不完备知识下的形式概念。第 3 节介绍了泛化粗糙集理论的基本概念。不完备背景下的形式概念表示与计算在第 4 节中定义和研究。一个简单地描述这种方法过程的例子在第 5 节。最后是结论。

2 形式概念与不完备知识

形式概念分析是反映概念的哲学理解的一种格理论形式和集理论模型。一个概念是由两个部分构成的一个思想单元：包含所有属于概念的对象的外延和包含所有对象共有属性的内涵。表现一个具体领域知识的基本模型是形式背景。一个形式背景由一个对象集、一个属性集和一个表明一个对象是否拥有一个属性的关系构成。在形式背景上存在一个反映子概念和超概念关系的自然序，所有这个序的集合就是概念格。

具有未知或缺失项的形式背景可以用所谓三值背景表示 $(G,M,\{\times,o,?\},I)$[6,13]。G 是对象集，M 是属性集，$I:G\times M\to\{(\times,o,?\}$ 即 $I(g,m\in\{\times,o,?\},\ g\in G,m\in M$。这里问号表示对象 g 是否有属性 m 是未知或缺失的。我们称其为不完备信息系统或背景。对两个拥有相同对象集和属性集不完备背景 $K_1=(G,M,\{\times,?,o\},I_1)$ 和 $K_2=(G,M,\{\times,?,o\},I_2)$，可以比较它们的信息。如果 K_2 能由 K_1 用 $\times$ 或 o 替换问号派生，那么背景 K_2 比背景 K_1 包含更多的信息，表为 $K_1\leqslant K_2$，也就是说 $\leqslant$ 是信息序。一个背景 K 的完备(completion)是一个源于 K 的、用非问号值替换问号值的完备背景(=形式背景)。因此，K 的完备是 K 上的(在信息序上的)最大背景。在本文中，我们使用 COMP 表示 K 的完备，COMP (K) 是 K 的所有完备的集。一个包含 n 个问号的不完备背景具有 2^n 个完备。

令 $K=(G,M,\{\times,?,o\},I)$ 是一个不完备背景，对 $B,B\subseteq M$ 的确定的外延 $B^{\square}$ 是所有确定拥有 B 的属性的对象集，$B^{\square}=\{g\in G|I(g,m)=\times$ 对所有 $m\in B\}=\bigcap_{m\in B}Im$，$Im=\{g\in G|I(g,m)=\times\}$[5]。$B$ 的可能的外延 $B^{\diamond}$ 是所有可能拥有 B 的属性的对象集，$B^{\diamond}=\{g\in G|I(g,m)\neq o$ 对所有 $m\in B\}=\bigcap_{m\in B}Im$，$Im=\{g\in G|I(g,m)\neq o\}$。$S,S\subseteq G$ 的确定内涵 $S^{\square}$ 和可能内涵 $S^{\diamond}$ 的定义与之类似：

$S^{\square}=\{m\in M|I(g,m)=\times$ 对所有 $g\in S\}=\bigcap_{g\in S}gI$，$gI=\{m\in M|I(g,m)=\times\}$

$S^{\diamond}=\{m\in M|I(g,m)\neq o$ 对所有 $g\in S\}=\bigcap_{g\in S}gI$，$gI=\{m\in M|I(g,m)\neq o\}$

如果 $K=(G,M,\{\times,?,o\},I)$ 是完备的(或是一个形式背景 $K=(G,M,I)$)，则有 $B^{\square}=B^{\diamond}:=B'$ 和 $S^{\square}=S^{\diamond}:=S'$。其正是文[1]中定义的外延和内涵。操作符 $':\wp(G)\to\wp(M)$ 和 $':\wp(M)\to\wp(G)$ 被称为演算，$\wp(G)$ 是 G 的幂集，$\wp(M)$ 是 M 的幂集。形式背景 (G,M,I) 上的一个形式概念是满足 $S\subseteq G,B\subseteq M,S'=B$ 和 $B'=S$ 的集对 (S,B)。S 是概念的外延，B 是概念的内涵。(G,M,I) 的所有形式概念的有序集就是 (G,M,I) 的概念格。形式化的格表示了背景中的所有概念以及与属性的关系。在一个给定的背景上，新概念可以通过格关系来发现，因此知识可以通过探索技术获取。

更一般地，$(G,M,\{\times,?,o\},I)$ 也被称为一个形式背景。

3 自反相似关系上的粗糙集

相似信息可以用每个对象 $x\in G$ 的相似类来表示。G 是一个对象集。更确切地说，x 的相似类 $R(x)$ 是与 x 相似的对象集：$R(x)=\{y\in G|\ yRx\}$。考虑 R 的逆关系 R^{-1}。令 $R^{-1}(x)$ 是 x 与之相似的对象的类 $R^{-1}(x)=\{y\in G|\ xRy\}$。注意这里的表述，$yRx$ 意味着 y 与 x 相似，其具有方向性。对一个子集 $X\subseteq G$ 和一个 G 上的二元关系 R，任何对象 $x\in G$ 属于且仅属于下列分类之一：

正对象：如果 $x\in X$ 并且 $R^{-1}(x)\subseteq X$，x 确切地属于 X；

一类不确切对象：如果 $x\in X$ 并且 $R^{-1}(x)\cap(G\backslash X)\neq\varnothing$，$x$ 是一类不确切对象；

二类不确切对象：如果 $x\in G\backslash X$ 并且 $R^{-1}(x)\cap X\neq\varnothing$，$x$ 是二类不确切对象；

负对象：如果 $x\in G\backslash X$ 并且 $R^{-1}(x)\subseteq G\backslash X$，$x$ 确切地不属于 X；

因此，四个分类确定了 G 的分割。

对象的不确定性自然引出了基于自反二元关系 R 的粗近似定义。

定义 1 对于 G 上的子集 $X\subseteq G$ 和二元自反关系 R，X 的下近似表为 $R_*(X)$，X 的上近似表为 $R^*(X)$。其定义如下[4]：$R_*(X)=\{x\in G|R^{-1}(x)\subseteq X\}$

$R^*(X)=\bigcup_{x\in X}R(x)=\{x\in G|R^{-1}(x)\cap X\neq\varnothing\}$

一个需要解决的问题是根据 R 提供的信息，一个 G 上的子集 $X\subseteq G$ 是否能够被正确地刻画。

定理 1 对于 G 上的子集 $X\subseteq G$ 和二元自反关系 R，如果 $R_*(X)=R^*(X)$，那么 X 是 R 可定义(以 X 为所指向的集)的[4]

如果 R 还是对称和传递的，可定义集即能表为 R 的等价类的并，称为 G 的可定义集。

定理 2 如果 $X\subseteq G$，$Y\subseteq G$ 是定义在 G 上的子集和 R 二元关系，则

1) $R_*(X)\subseteq X\subseteq R^*(X)$

2) $X\subseteq Y\Rightarrow R_*(X)\subseteq R_*(Y)$

3) $X\subseteq Y\Rightarrow R^*(X)\subseteq R^*(Y)$

4) $R_*(X\cap Y)=R_*(X)\cap R_*(Y)$

5) $R^*(X\cup Y)=R^*(X)\cup R^*(Y)$

6) $R_*(X\cup Y)\supseteq R_*(X)\cup R_*(Y)$

7) $R^*(X\cap Y)\subseteq R^*(X)\cap R^*(Y)$

证明：仅证 2),3),4),6),7)。

2) $x\in R_*(X)\Rightarrow R^{-1}(x)\subseteq X,X\subseteq Y\Rightarrow R^{-1}(x)\subseteq Y\Rightarrow x\in R_*(Y)\Rightarrow R_*(X)\subseteq R_*(Y)$。

3) $X\subseteq Y\Rightarrow\bigcup_{x\in X}R(X)\subseteq\bigcup_{x\in Y}R(x)\Rightarrow R^*(X)\subseteq R^*(Y)$。

4) $x\in R_*(X\cap Y)\Leftrightarrow R^{-1}(x)\subseteq X\cap Y\Leftrightarrow R^{-1}(x)\subseteq X$ 并且 $R^{-1}(x)\subseteq Y\Leftrightarrow x\in R_*(X)$ 并且 $x\in R_*(Y)\Leftrightarrow x\in R_*(X)\cap R_*(Y)$。

6) 因为 $X\subseteq X\cup Y$，$Y\subseteq X\cup Y$，从 2) 即可得到 $R_*(X)\subseteq R_*(X\cup Y)$ 并且 $R_*(Y)\subseteq R_*(X\cup Y)$。故，$R_*(X\cup Y)\supseteq R_*(X)\cup R_*(Y)$。

7) $X\cap Y\subseteq X,X\cap Y\subseteq Y\Rightarrow R^*(X\cap Y)\subseteq R^*(X)$ 并且 $R^*(X\cap Y)\subseteq R^*(Y)\Rightarrow R^*(X\cap Y)\subseteq R^*(X)\cap R^*(Y)$。

4 概念可定义性与不完备背景上的概念表示

给定一个形式背景(G, M, I)，一个属性$m\in M$被称为可定义属性，当它的外延$Im\subseteq G$是关于G上不可分辨关系对象的一个可定义子集。一个可定义形式背景，其所有属性都是可定义的。对任意形式背景$(G, M, \{\times, o, ?\}, I)$，我们希望用可定义背景来近似$I$。使用两个类似于文[5]中的概念：可能的上近似和必然的下近似。这两个背景上的近似提供了$(G, M, \{\times, o, ?\}, I)$上概念的上下近似表示。

考虑Im的可能和确定的外延，问题可以分两部分讨论。

4.1 $I(g,m)\neq o$. Im是可能拥有属性m的所有对象的集

定义2 令$(G, M, \{\times, o, ?\}, I)$是一个形式背景，$Im=\{g\in G|I(g,m)\neq o\}$，$gI=\{m\in M|I(g,m)\neq o\}$。考虑$G$上的二元自反关系$R$，$R(x)=\{y\in G|\ I(x,m)=I(y,m)\text{or } I(x,m)=?\}$，$R^{-1}(x)=\{y\in G|\ I(x,m)=I(y,m)\text{or } I(y,m)=?\}$。$I$的$R$上近似表为$I^{*R}$：对每个属性$m\in M$，$m$在上近似$I^{*R}$的外延是其在$I$的外延的上近似$I^{*R}m\overset{def}{=}(Im)^{*R}=\bigcup_{x\in Im}R(x)=\{x\in G|R^{-1}(x)\cap Im\neq\varnothing\}$。$I$的$R$下近似表为$I_{*R}$：对每个属性$m\in M$，$m$在下近似$I_{*R}$的外延是其在$I$的外延的下近似$I_{*R}m\overset{def}{=}(Im)_{*R}=\{x|R^{-1}(x)\subseteq Im\}$。

这里x与y是等价关系，其一定满足二元自反关系，反之则不成立。也就是说，等价关系包含在自反二元关系中。

定理3 令$(G, M, \{\times, o, ?\}, I)$是一个形式背景，$B\subseteq M$。在$I$中的$B$的外延的上下近似包含在$I$的$R$上下近似的外延中。

证明：假设$B'_{I^{*R}}$是B在I^{*R}中的外延并且$B'_{I_{*R}}$是B在I_{*R}中的外延，则有$B'_{I^{*R}}=\bigcap_{m\in B}I^{*R}m=\bigcap_{m\in B}(Im)^{*R}$。根据定理2.(7)，有

$$(\bigcap_{m\in B}Im)^{*R}\subseteq\bigcap_{m\in B}(Im)^{*R}$$

$$(B'_I)^{*R}=(\bigcap_{m\in B}Im)^{*R}\subseteq\bigcap_{m\in B}(Im)^{*R}=B'_{I^{*R}}$$

据定理2.(4)，得到

$$(B'_I)_{*R}=(\bigcap_{m\in B}Im)_{*R}=\bigcap_{m\in B}(Im)_{*R}=\bigcap_{m\in B}I_{*R}m=B'_{I_{*R}}$$

对于一个给定的背景$(G,M,\{\times,o,?\},I)$，当它的外延$A\subseteq G$是关于不可分辨关系对象的一个可定义子集时，形式概念(A, B)是一个可定义概念。

定理4 一个可定义形式背景的所有概念是可定义形式概念[5]。

定义3 一个可定义概念(A, B)的R-上近似是形式背景(G, M, I^{*R})上的可定义概念，被定义为$(A, B)^{*R}=(B'_{I^{*R}}, B''_{I^{*R}})$，$B''_{I^{*R}}=(B'_{I^{*R}})'$。$(A, B)$的形式背景$(G, M, I_{*R})$上$R$-下近似被定义为

$(A, B)_{*R}=(B'_{I_{*R}}, B''_{I_{*R}})$，$B''_{I_{*R}}=(B'_{I_{*R}})'$。

也就是说，一个概念可以由它的属性外延的可定义性来确定。

定理5 对$(G,M,\{\times,o,?\},I)$上的一个可定义概念(A, B)，且I_{*R}与I^{*R}不重叠（至少有一个$I_{*R}m\neq I^{*R}m, m\in B$）。如果$(A, B)_{*R}=(A, B)^{*R}$，则$(A, B)_{*R}$一定是背景上的一个完备的形式概念。

证明：$(A, B)_{*R}=(A, B)^{*R}$即$B'_{I^{*R}}=B'_{I_{*R}}$，$B''_{I^{*R}}=B''_{I_{*R}}$。存在$\bigcap_{m\in B}I_{*R}=\bigcap_{m\in B}I^{*R}\Rightarrow\bigcap_{m\in B}\{x\in G|R^{-1}(x)\cap Im\neq\varnothing\}=\bigcap_{m\in B}\{x\in G|R^{-1}(x)\subseteq Im\}\Rightarrow$对$m\in B$均有$R^{-1}(x)\subseteq Im$。由$Im$的取值知，$R(x)$是一个完备上的等价关系。其一定在COMP($K$)中。结论正确。

定义4 给定义概念(A, B)及其背景$(G,M,\{\times,o,?\}, I)$。S表示概念近似的精度。

$$S=\frac{1}{2}\left(\frac{|B'_{I_{*R}}\cap B'_{I^{*R}}|}{|B'_{I_{*R}}\cup B'_{I^{*R}}|}+\frac{|B''_{I_{*R}}\cap B''_{I^{*R}}|}{|B''_{I_{*R}}\cup B''_{I^{*R}}|}\right)$$，$|\cdot|$表示集合的势。

定理6 对一个背景$(G,M,\{\times,o,?\},I)$上的可定义概念(A, B)，下列结果成立：

a)$0\leqslant S\leqslant 1$。

b)$(A, B)_{*R}=(A, B)^{*R}\Rightarrow S=1$。

证明：a)根据前面的证明应有$(A, B)_{*R}\leqslant(A, B)^{*R}$。故

$$B'_{I^{*R}}\supseteq B'_{I_{*R}}, B''_{I^{*R}}\subseteq B''_{I_{*R}}\Rightarrow$$

$$|B'_{I^{*R}}\cap B'_{I_{*R}}|=|B'_{I_{*R}}|\leqslant|B'_{I^{*R}}\cup B'_{I_{*R}}|=|B'_{I^{*R}}|,$$

$$|B''_{I^{*R}}\cap B''_{I_{*R}}|=|B''_{I^{*R}}|\leqslant|B''_{I^{*R}}\cup B''_{I_{*R}}|=|B''_{I_{*R}}|$$

$$\Rightarrow S=\frac{1}{2}\left(\frac{|B'_{I_{*R}}|}{|B'_{I^{*R}}|}+\frac{B''_{I^{*R}}}{B''_{I_{*R}}|}\right)\leqslant 1。$$

b)$(A, B)_{*R}=(A,B)^{*R}\Rightarrow B'_{I^{*R}}=B'_{I_{*R}}, B''_{I^{*R}}=B''_{I_{*R}}\Rightarrow$

$$|B'_{I^{*R}}\cap B'_{I_{*R}}|=|B'_{I^{*R}}\cup B'_{I_{*R}}|,$$

$$|B''_{I^{*R}}\cap B''_{I_{*R}}|=|B''_{I^{*R}}\cup B''_{I_{*R}}|\Rightarrow S=1.\ \square$$

定理7 若(A, B)为背景$(G,M,\{\times,o,?\},I)$上的可定义概念，则$B'_{I_{*R}}\subseteq B'_{I^{*R}}, B''_{I_{*R}}=B''_{I^{*R}}$，$S$可简化为$S=\frac{|B'_{I_{*R}}|}{|B'_{I^{*R}}|}$。

证明：因为(A, B)为背景$(G,M,\{\times,o,?\},I)$上的可定义概念，有$(A, B)^{*R}=(B'_{I^{*R}}, B''_{I^{*R}})$，$(A, B)_{*R}=(B'_{I_{*R}}, B''_{I^{*R}})$，$B'_{I^{*R}}=\bigcap_{m\in B}I^{*R}m$，$B'_{I_{*R}}=\bigcap_{m\in B}I_{*R}m$。由定义知，对于每一个$m\in B$均有$I_{*R}m\subseteq I^{*R}m$，则必有$B'_{I_{*R}}=\bigcap_{m\in B}I_{*R}m\subseteq\bigcap_{m\in B}I^{*R}m=B'_{I^{*R}}$。

另外，考虑背景的上下近似的分类，

$$x,y\in G,\begin{cases}[x]_{I_{*R}}=\{y\in G|xI_{*R}=yI_{*R}\},\\ [x]_{I^{*R}}=\{y\in G|xI^{*R}=yI^{*R}\}\end{cases}$$。由$m\in B, I_{*R}m\subseteq I^{*R}m$可以推出，对象在$B'_{I_{*R}}$中的分类均为$B'_{I^{*R}}$中的子类。也就是说，$I_{*R}$比$I^{*R}$分类更细。故

$$B''_{I_{*R}}=\bigcap_{g\in B'_{I_{*R}}}gI_{*R}=\bigcap_{g\in B'_{I^{*R}}}gI^{*R}=B''_{I^{*R}}$$

直接由定义4可得到$S=\frac{1}{2}\left(1+\frac{|B'_{I_{*R}}|}{|B'_{I^{*R}}|}\right)$。由于相似度只受外延的影响，因此可直接计算$S=\frac{|B'_{I_{*R}}|}{|B'_{I^{*R}}|}$。

推论：I^{*R}上概念的个数$\leqslant I_{*R}$上概念的个数。

定理8 对$(G,M,\{\times,o,?\},I)$上的一个可定义概念(A, B)，以下结论正确：$(A, B)_{*R}\leqslant(A, B)^{*R}$，即$B'_{I_{*R}}\subseteq B'_{I^{*R}}$ and $B''_{I^{*R}}\subseteq B''_{I_{*R}}$。

证明：由定理7易知结论成立。

定理8保证了定义3的合理性。

定理9 背景的R-下近似是一个完备，即R-下近似可以从背景中将所有“?”取“×”得到。

证明：因为$I_{*R}m=\{x\in G|R^{-1}(x)\subseteq Im\}$。对于每一个$Im$，若$I(x,m)=\times$，则$I(y,m)=\times$或$I(y,m)=?$，$R^{-1}(x)\subseteq Im$成立；若$I(x,m)=?$，$R^{-1}(x)\subseteq Im$成立，则$I(y,m)=?$；若$I(x,m)=o$，则$x\notin I_{*R}m$。所以，由$Im$取值知，$I_{*R}$即由原

背景中所有"?"取"×"得到。

以下是一个计算近似概念的算法。

算法

1. 输入背景$(G,M,\{\times,o,?\},I)$
2. 计算 $R(x)$, $R^{-1}(x)$
3. 计算 $I^{*R}m=\bigcup_{x\in Im}R(x)$,
   $I_{*R}m=\{x\in G|R^{-1}(x)\subseteq Im\}$
4. 输出 (A, B)
5. $B'_{I^{*R}}\leftarrow\bigcap_{m\in B}I^{*R}m, B''_{I^{*R}}\leftarrow\bigcap_{g\in B'_{I^{*R}}}gI$
   $B'_{I_{*R}}\leftarrow\bigcap_{m\in B}I_{*R}m, B''_{I_{*R}}\leftarrow\bigcap_{g\in B'_{I_{*R}}}gI$
6. If $(A, B)_{*R}\leqslant(A, B)^{*R}$
   $(A, B)_{*R}\leftarrow(B'_{I_{*R}}, B''_{I_{*R}})$, $(A, B)^{*R}\leftarrow(B'_{I^{*R}}, B''_{I^{*R}})$
   else
   (A, B)不可近似.
7. $S\leftarrow\dfrac{|B'_{I_{*R}}|}{|B'_{I^{*R}}|}$
8. end

值得注意的是,以上提到的$R(x)$是自反的。如果$R(x)$是容错相似关系,结论也是正确的。事实上,$R(x)=\{y\in G|I(x,m)=I(y,m)$or $I(x,m)=?$ or $I(y,m)=?\}$, $R^{-1}(x)=R(x)$。

$I_{*R}m=\{x|R(x)\subseteq Im\}$, $I^{*R}m=\{x\in|R(x)\cap Im\neq\varnothing\}$。如果$R(x)$是一个等价关系,$R(x)=\{y\in G|I(x,m)=I(y,m)\}$。在这种情况下,背景是一个多值背景,可以如文[1,9]那样处理。

4.2 $I(g,m)=\times$. Im 是确定拥有属性 m 的所有对象的集

此时,所有的问号"?"都被看作是"o",背景是二元的,概念的表示可以用 Kent 方法[5]。

5 一个例子

为说明本文介绍的方法,用表 1,2 给出的不完备背景演示了一个实例。这个背景是对房屋情况的一个简单描述。这里用"L", "B" 和"F"表示"Location", "Basement" 和 "Fireplace", "Good","Bad" 和 "Yes" ,"No"对应"×" 和 "o"。接下来我们将进行这个房屋背景的粗概念分析。

表 1

I	Location	Basement	Fireplace
1	Good	No	Yes
2	Bad	?	No
3	Good	No	?
4	Bad	Yes	No
5	?	?	Yes

表 2

I	L	B	F
1	×	o	×
2	o	?	o
3	×	o	?
4	o	×	o
5	?	?	×

$R(1)=\{1\}$, $R^{-1}(1)=\{1,3,5\}$, $R(2)=\{2,4\}$, $R^{-1}(2)=\{2\}$, $R(3)=\{1,3\}$, $R^{-1}(3)=\{3\}$, $R(4)=\{4\}$, $R^{-1}(4)=\{2,4\}$, $R(5)=\{1,5\}$, $R^{-1}(5)=\{5\}$。

$I_L=\{1,3,5\}$, $I_B=\{2,4,5\}$, $I_F=\{1,3,5\}$。

表 3,4 表示了 R 关系的上下似背景。背景的 3 个完备 $COMP_1$, $COMP_2$ 和 $COMP_3$ 在表 4 中。注意到属性值有 4 个是标有问号的,因此背景存在 16 个完备。

表 3

I^{*R}	L	B	F
1	×	×	×
2	o	×	o
3	×	o	×
4	o	×	o
5	×	×	×

I_{*R}	L	B	F
1	×	o	×
2	o	×	o
3	×	o	×
4	o	×	o
5	×	×	×

上近似背景的形式概念是(15, LBF), (1245, B)和(135, LF)。下近似背景中的是(135, LF), (245, B)和(5, LBF)。

表 4 $COPM_1$, $COPM_2$ 和 $COPM_3$

I	L	B	F
1	×	o	×
2	o	o	o
3	×	o	o
4	o	×	o
5	o	o	×

I	L	B	F
1	×	o	×
2	o	×	o
3	×	o	o
4	o	×	o
5	o	×	×

I	L	B	F
1	×	o	×
2	o	o	o
3	×	o	×
4	o	×	o
5	×	o	×

$COMP_1$ 中概念的上下近似是$(13,L)^*=(135, LF)$, $(13, L)_*=(135, LF)$, $S=1$; $(1, LF)^*=(135, LF)$, $(1, LF)_*=(135, LF)$, $S=1$; $(4,B)^*=(1245,B)$, $(4,B)_*=(245,B)$, $S=7/8$; $(15,F)^*=(135, LF)$, $(15,F)_*=(135, LF)$, $S=1$。$COMP_2$ 概念的上下近似是$(13,L)^*=(135, LF)$, $(13,L)_*=(135, LF)$, $S=1$; $(245,B)^*=(1245,B)$, $(245,B)_*=(245,B)$, $S=7/8$; $(15,F)^*=(135, LF)$, $(15, F)_*=(135, LF)$, $S=1$; $(5, BF)^*=(15, LBF)$, $(5, BF)_*=(5, LBF)$. $S=3/4$。

很明显,(135, LF)是 $COPM_3$ 的一个形式概念。

结论 本文提出了一个新的不完备知识下的概念近似方法。这个方法是基于泛化粗糙集理论的。这种方法可以将对可能的众多情况的考虑,简化为对两个确定的形式背景的研究。这对大背景下、多未知或缺失值的形式概念分析研究十分有益。

参 考 文 献

1 Ganter B, Wille R. Formal Concept Analysis. In: Mathematical Foundations, Berlin: Springer, 1999
2 Pawlak Z. Rough Sets. International Journal of Computer and Information Sciences, 1982, 11: 341～356
3 Pawlak Z. Rough Sets, Theoretical Aspects of Reasoning About Data. Dordrecht: Kluwer Academic Publishers, 1991
4 Slowinski R, Vanderpooten D. A generalized definition of rough approximations based on similarity. IEEE Transactions on Data and Knowledge Engineering, 2000, 12(2): 331～336
5 Kent R. Rough concept analysis: A synthesis of rough sets and formal concept analysis. Funadamenta Informaticae, 1996, 27: 169～181
6 Holzer R. Knowledge acquisition under incomplete knowledge using methods from formal concept analysis Part I, Part II. Fundamenta Informaticae, 2004, 63(1)

(下转第 190 页)

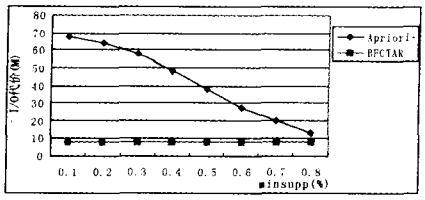

图 5 BFCTAR 算法和 Apriori+算法 I/O 代价比较

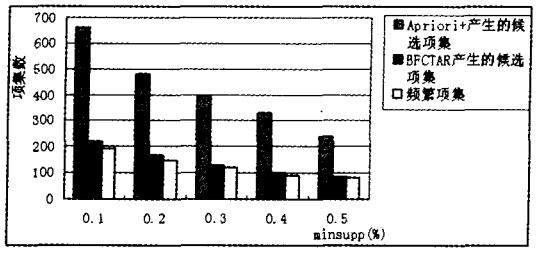

图 6 BFCTAR 算法和 Apriori+算法生成的候选项集数比较

BFCTAR 算法与 Apriori+算法的比较结果如图 4,5,6 所示。可看出,基于模糊日历约束的 BFCTAR 算法的执行效率优于 Apriori+算法,特别是在支持度阈值较小时,其 I/O 代价也明显小于 Apriori+算法;其产生的候选项集数接近于频繁项集数,明显小于 Apriori+算法所产生的候选项集数。其原因主要在于 BFCTAR 算法在生成候选项集的过程中采用累进的方法,每扫描一个分区即记录候选项集的相应的信息,减少了扫描数据库的次数,并且在处理过程中不断删除不满足条件的项集,所以生成的候选项集数也很接近于频繁项集。

小结 本文在模糊集理论和模糊日历代数的基础上,提出了一种基于模糊日历约束的时序关联规则挖掘算法,理论分析和实验结果表明该算法是有效的。虽然是基于日历约束的模糊规则挖掘算法,但对于其它属性的模糊约束也具有普遍意义。

参 考 文 献

1 Han J, Dong G, Yin Y. Efficient Mining of Partial Periodic Patterns in Time Series Databases. In: Proceedings of the International Conference on Data Engineering, 1999. 106～115
2 Ozden B, Ramaswamy S, Silberscbatz A. Cyclic Association Rules. In: Proceedings of the 15th International Conference on Data Engineering, 1998. 412～421
3 Li Y, Ning P, Wang X S, Jajodia S. Discovering Calendar-based Temporal Association Rules. Data and Knowledge Engineering, 2003, 44(2): 193～218
4 Ramaswamy S, Mahajan S, Silberschatz A. On the Discovery of Interesting Patterns in Association Rules. In: Proceedings of the International Very Large Database Conference, 1998. 368～379
5 Lee W J, Jiang J Y, Lee S J. An Efficient Algorithm to Discover Calendar- based Temporal Association Rules. In: Proceedings of 2004 IEEE International Conference on Systems, Man, and Cybernetics, 2004. 3122～3127
6 Lee W J, Lee S J. Fuzzy Calendar Algebra and Its Applications to Data Mining. In: Proceedings of 11th International Symposium on Temporal Representation and Reasoning, 2004. 71～78
7 Lee C H, Lin C R, Chen M S. Sliding-window filtering: An efficient algorithm for incremental mining. In: Proc. ACM 10th Int. Conf. Information Knowledge Management, Atlanta, GA, Nov. 2001. 263～270
8 Lee C H, Ou J C, Chen M S. Progressive weighted miner: An efficient method for time-constraint mining. In: Proc. 7th Pacific-Asia Conf. Knowledge Discovery Data Mining, Seoul, Korea, Apr.-May 2003. 449～460
9 Cheung D W, Lee S D, Kao B. A general incremental technique for maintaining discovered association rules. In: Proc. 5th Int. Conf. Database Systems Advanced Applications, Melbourne, Australia, Apr. 1997. 185～194
10 Ale J M, Rossi G H. An Approach to Discovering Temporal Association Rules. In: Proc. of the 2000 ACM Symposium on Applied Computing, 2000. 294～300
11 Agrawal R, Srikant R. Fast algorithms for mining association rules. In: Proc. 1994 Int. Conf. Very Large Databases (VLDB'94)

(上接第 169 页)

7 Pensa R G, Boulicaut J-F. Towards Fault-Tolerant Formal Concept Analysis. In: the 9th Congress of the Italian Association for Artificial Intelligence, Milan, Italy, September Springer-Verlag LNAI 3673, 2005. 212～223
8 Grzymala-Busse J W. Three Approaches to Missing Attribute Values-A Rough Set Perspective. In: Workshop on Foundations of Data Mining, associated with the fourth IEEE International Conference on Data Mining, Brighton, UK, November, 2004, 1: 4
9 Saquer J, Deogun J. Concept approximations based on rough sets and similarity measures Int [J]. Appl Math and Comp Sci, 2001, 11(3): 655～674
10 Grzymala-Busse J W, Siddhaye S. Rough Set Approaches to Rule Induction from Incomplete Data. In: Proceedings of the IPMU' 2004, the 10th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, Perugia, Italy, July 2004. 923～930
11 Grzymala-Busse J W. Rough Set Strategies to Data with Missing Attribute Values. In: Proceedings of the Workshop on Foundations and New Directions in Data Mining, associated with the third IEEE International Conference on Data Mining, Melbourne, FL, USA, November 2003. 56～63
12 Kryszkiewicz M. Rules in incomplete information systems. Information Sciences: an International Journal, 1999, 113(3-4): 271～292
13 Burmeister P, Holzer R. Treating Incomplete Knowledge in Formal Concept Analysis. In: Ganter B, Stumme G, Wille R, eds. Formal Concept Analysis: State of the Art. LNAI 3626. Springer, Heidelberg, 2005
14 Wang Guo-yin. Extension of rough set under incomplete information systems. Journal of Computer Research and Development, 2002, 39(10): 1238～1243
15 Priss U. Formal concept analysis in information science. Annual Review of Information Science and Technology, 2006, 40: 521～543
16 Grzymala-Busse J W. Data with missing attribute values: Generalization of indiscernibility relation and rule induction. Transactions on Rough Sets, Lecture Notes in Computer Science Journal Subline, Springer-Verlag, 2004, 1: 78～95
17 Kryszkiewicz M. Rough set approach to incomplete information systems. In: Proceedings of the Second Annual Joint Conference on Information Sciences, September 28-October 1, Wrightsville Beach, NC, 1995. 194～197
18 Stefanowski J. Algorithms of Decision Rule Induction in Data Mining. Poznan, Poland: Poznan University of Technology Press, 2001
19 Stefanowski J, Tsoukias A. On the extension of rough sets under incomplete information. In: Proceedings of the 7th International Workshop on New Directions in Rough Sets, Data Mining, and Granular-Soft Computing, RSFDGrC'1999, Yamaguchi, Japan, 1999. 73～81
20 Stefanowski J, Tsoukias A. Incomplete information tables and rough classification. Computational Intelligence, 2001, 17: 545～566
21 Greco S, Matarazzo B, Slowinski R. Dealing with missing data in rough set analysis of multi-attribute and multicriteria decision problems. In: Zanakis S. H, Doukidis G, Zopounidis Z, eds. Decision Making: Recent developments and Worldwide Applications, Dordrecht, Boston, London: Kluwer Academic Publishers, 2000. 295～316
22 Slowinski R, Vanderpooten D. A generalized definition of rough approximations based on similarity. IEEE Transactions on Knowledge and Data Engineering, 2000, 12: 331～336
23 Yao Y Y. Two views of the theory of rough sets in finite universes. International J of Approximate Reasoning, 1996, 15: 291～317
24 Yao Y Y. Relational interpretations of neighborhood operators and rough set approximation operators. Information Sciences, 1998, 111: 239～259
25 Yao Y Y. On the generalizing rough set theory. In: Proc. of the 9th Int Conference on Rough Sets, Fuzzy Sets, Data Mining and Granular Computing (RSFDGrC' 2003), Chongqing, China, Oct. 2003. 44～51