文章编号:1001 - 9081(2006)11 - 2664 - 03

# 基于集合覆盖的不完备信息系统属性约简方法

冯朝一[1],梁家荣[1],黄柳萍[1],李天志[2]
(1. 广西大学 计算机与电子信息工程学院,广西 南宁 530004; 2. 德州学院 计算机系,山东 德州 253023)
(fcy947@sohu.com)

**摘 要**:深入分析了不完备信息系统的特点以及相似关系的定义,通过构造不完备信息系统的相关矩阵,把不完备信息系统的最小属性约简问题与最小集合覆盖问题联系起来,将不完备信息系统的最小属性约简问题转化为最小集合覆盖问题,给出了基于集合覆盖的不完备信息系统最小属性约简算法。实例分析证明该算法可行,高效。

**关键词**:集合覆盖;最小属性约简;粗糙集;相关矩阵
**中图分类号**: TP311.131 **文献标识码**:A

# Attribute reduction way of incomplete information system based on set covering problem

FENG Chao-yi[1], LIANG Jia-rong[1], HUANG Liu-ping[1], LI Tian-zhi[2]
(1. *College of Computer and Electronic Information, Guangxi University, Nanning Guangxi* 530004, *China*;
2. *Computer Department, Dezhou University, Dezhou Shandong* 253023, *China*)

**Abstract:** By analyzing the characteristics of incomplete information system and the definition of similar relation, and constructing the related matrix of incomplete information system, the minimum attribute reduction problem was related to the minimum set covering problem. The minimum attribute reduction problem could be translated to the set covering problem, and the minimum attribute reduction could be got by using the set covering problem ways. The examples prove that this method is feasible and efficient.

**Key words:** set covering; minimum attribute reduction; rough set; relation matrix

## 0 引言

粗糙集理论是一种新的处理模糊和不确定性知识的数学工具[1]。其主要思想就是在保持分类能力不变的前提下,通过知识约简,导出问题的决策或分类规则[1]。属性约简是其核心内容之一,即在导出决策规则前约去冗余属性,得到最佳决策规则。求所有的约简或相对约简已经被证明是 NP 完全问题[2],故一般采用启发式信息来求属性约简。针对完备信息系统的最小属性约简方法有基于正区域的方法[3~5]、基于信息熵[6]的方法、基于区分矩阵属性的方法[7]等。

关于不完备信息系统的属性约简的研究还不够成熟,文献[8,9]基于信息熵的不完备信息系统的属性约简算法能够求得约简,但不一定能得到最小约简,其算法复杂度为 $O(|A|^3|U|^2)$。本文通过分析不完备信息系统的定义以及不完备信息系统相似关系的定义,利用不完备信息系统的相容类给出构造相关矩阵的方法,把不完备信息系统的最小属性约简问题与求集合的最小覆盖问题联系起来,求出各属性的区分集合,进而求出各相容类的最小集合覆盖即最小属性约简,给出了不完备信息系统最小属性约简算法。理论分析和实验结果表明,该算法时间复杂度为 $O(|A|^2|U|^2)$,在效率上有较大提高。

## 1 粗糙集的相关知识[2]

### 1.1 完备信息系统

$S=(U,A,V,f)$ 是信息系统,其中:$U$ 为对象的非空有限集合;$A$ 为属性的非空有限集合;$V=\bigcup_{a\in A}V_a$,$V_a$ 是属性 $a$ 的值域;$f:U\times A\to V$ 是一个信息函数,它为每个对象的每个属性赋予一个信息值,即 $\forall a\in A, x\in U, f(x,a)\in V_a$。

### 1.2 不可区分关系

属性子集 $P\subseteq A$ 决定了一个不可区分关系 $ind(P)$:$Ind(P)=\{(x,y)\in U\times U \mid \forall a\in P, a(x)=a(y)\}$。

### 1.3 不完备信息系统

对于一个信息系统 $S=(U,AT,V,f)$,一些属性值可能是缺省的,为了表明这种情况,通常给定一个区分值(即空值)给这些属性。如果至少有一个属性 $a\in AT$ 使得 $V_a$ 含有空值,则称 $S$ 为一个不完备信息系统,否则它是完备的,我们用“*”表示空值。

### 1.4 相似关系

令 $A\subseteq AT$,我们定义相似关系如下:$SIM(A)=\{(x,y)\in U\times U \mid \forall a\in A, a(x)=a(y) \text{ or } a(x)=* \text{ or } a(y)=*\}$。

### 1.5 相容类

**定理** $SIM(A)$ 是一个相似关系:$SIM(A)=\bigcap_{a\in A}SIM(\{a\})$。

**收稿日期**:2006 - 05 - 16;**修订日期**:2006 - 07 - 23 **基金项目**:教育部留学回国人员科研专项基金资助项目(教外司留[2004]527);广西高校百名学科带头人专项基金资助项目(桂人教[2003]97 号);广西研究生教育创新计划项目

**作者简介**:冯朝一(1975 - ),男,河南驻马店人,硕士研究生,主要研究方向:数据挖掘、人工智能; 梁家荣(1966 - ),男,广西玉林人,教授,主要研究方向:数据挖掘、人工智能.

令 $S_A(x)$ 表示对象集$\{y \in U \mid (x,y) \in SIM(A)\}$,对于$A$而言,$S_A(x)$ 是与 $x$ 可能不可区分的对象的最大集合。

令 $D_A(x)$ 表示对象集$\{y \in U \mid (x,y) \notin SIM(A)\}$, 对于$A$而言,$D_A(x)$ 是与 $x$ 可能可区分的对象的最大集合。

对任意 $x \in U$, $S_A(x) \cap D_A(x) = \varnothing$ 且 $S_A(x) \cup D_A(x) = U$。

令 $U/SIM(A) = \{S_A(x) \mid x \in U\}$ 表示分类。$U/SIM(A)$ 中的任何元素称为相容类。$U/SIM(A)$ 中的相容类一般不构成 $U$ 的划分,它们构成 $U$ 的覆盖, $\cup U/SIM(A) = U$。

### 1.6　属性约简

信息系统中属性并不是同等重要的,甚至其中某些属性是冗余的。属性约简就是在保持知识库分类能力不变的前提下,删除其中不相关或不重要的知识。

形式上,一个集合 $A \subseteq AT$ 是信息系统的一个约简,若 $SIM(A) = SIM(AT)$ 且 $\forall B \subset A, SIM(B) \neq SIM(AT)$。一般情况下信息系统的属性约简是不唯一的,人们希望得到最小约简,即具有最少属性的约简。在不完备信息系统中,就是从 $AT$ 中找出个数最少的属性集 $A$,使得 $SIM(A) = SIM(AT)$。

## 2　最小集合覆盖问题[4]

**定义**　$S$ 是一个集合,$S_1, S_2, \cdots, S_m$ 是 $S$ 的子集,且构成 $S$ 的覆盖,即 $\bigcup_{i=1}^{m} S_i = S$,求最小的覆盖。

由定义可以知道求最小集合覆盖的前提是集合的所有子集能构成集合的覆盖。

## 3　不完备信息系统的相关矩阵

由相似关系 $SIM(A)$ 的定义可以知道,在同一个相容类中的所有对象都满足 $\forall a \in A, a(x) = a(y)$ or $a(x) = *$ or $a(y) = *$,即不能区分同一相容类中的任意两个对象。同理我们知道,在两个不同相容类中的所有对象间至少存在一对对象满足 $\exists a \in A, a(x) \neq a(y)$,即在不同的相容类中至少存在一个属性使得两个相容类的至少一对对象在此属性下的值不同,也就是说 $a$ 是两个不同相容类的区别属性。区别属性是维持 $A \subseteq AT, SIM(A) = SIM(AT)$ 的关键,如果存在最小的区别属性 $A$,使得 $SIM(A) = SIM(AT)$,则 $A$ 就是 $AT$ 的最小约简。

**定义 1**　不完备信息系统 $S = (U, AT, V, f)$ 的相关矩阵

假定不完备信息系统 $S = (U, AT, V, f)$ 的相容类簇 $X = \{X_1, X_2, \cdots, X_r\}$ 和属性集 $AT = \{a_1, a_2, \cdots, a_n\}$,由相容类簇 $X$ 中任意两个不同相容类组成集合记为:$U_1 = \{(X_i, X_j) \mid X_i, X_j \in X, i < j \leqslant r\}$,显然 $U_1$ 中对象的个数 $|U_1| = r*(r-1)/2$。定义一个 $r*(r-1)/2$ 行 $n$ 列的矩阵 $\boldsymbol{M}$,其中行代表 $U_1$ 中的对象,列代表 $AT$ 中的元素。矩阵第 $i$ 行第 $j$ 列元素 $m_{ij}$ 取值如下:

任给两个不同的相容类 $X_m, X_n$;$u_i = (X_m, X_n) \in U_1$, $a \in AT$,有:

$$m_{ij} = \begin{cases} 1, & a(x) \neq a(y), \exists (x,y) \in X_m \times X_n \\ 0, & a(x) = a(y) \cdot \text{or} \cdot a(x) = * \cdot \text{or} \cdot a(y) = *, \\ & \forall (x,y) \in X_m \times X_n \end{cases}$$

由这样的 $m_{ij}$ 组成的矩阵 $\boldsymbol{M} = (m_{ij})$,称为不完备信息系统的相关矩阵。

不完备信息系统的相关矩阵具有以下几个定理。

**定理 1**　相关矩阵中没有全零的行向量。

证明:　根据相似关系的定义,任意两个不同相容类 $X_m, X_n$,之间一定存在区分属性,即$\{\exists (x,y) \in X_m, X_n\} \mid \exists a \in AT, a(x) \neq a(y)\}$ 否则 $X_m, X_n$ 中的元素满足相似关系,属于同一相容类,证毕。

**定理 2**　在相关矩阵 $\boldsymbol{M}$ 中,如果 $m_{ij} = 1$,则属性 $a$ 是 $u_i = (X_m, X_n) \in U_1$ 对应的两个相容类的区分属性,否则 $a$ 不是 $u_i = (X_m, X_n) \in U_1$ 对应的两个相容类的区分属性。

证明:　根据相关矩阵的定义如果 $m_{ij} = 1$,则 $\exists (x,y) \in X_m \times X_n$,使得 $a(x) \neq a(y)$, 即属性 $a$ 可以区分两个相容类 $X_m, X_n$,反之则属性 $a$ 不能区分两个相容类 $X_m, X_n$,证毕。

**定理 3**　在相关矩阵中如果存在全零的列,则相应属性直接约去,而得到一个约简,但不一定是最小约简。

证明:　根据区分矩阵的定义,该属性不能区分任何相容类,对维持 $SIM(A) = SIM(AT)$ 没有任何贡献,即可以约去,证毕。

**定义 2**　区分集合$\{u_i \mid m_{ij} = 1, u_i = (X_m, X_n) \in U_1\}$

由 $U_1$ 中所有能被 $a_i$ 区分的 $u = (X_m, X_n) \in U_1$ 的全体组成的集合叫作属性 $a_i$ 的区分集合,记为 $S_i$。可以表示为:

$$S_i = \bigcup_{i=1}^{|U_1|} \{u_i \mid m_{ij} = 1, u_i = (X_m, X_n) \in U_1\}$$

这样就可以把每个属性的区分集合看作是 $U_1$ 的子集,实际上所有的属性的区分集合能够覆盖 $U_1$。即有如下定理。

**定理4**　所有属性的区分集合构成集合的覆盖,即 $U_1 = \bigcup_{i=1}^{n} S_i$。

证明:　由相关矩阵的定义和定理 1 我们知道集合 $U_1$ 中不存在全零的行向量,再由区分集合的定义可以知道对集合 $U_1$ 中的任意元素它必定属于某一属性的区分集合,这样它必定属于所有区分集合的并集,即定理成立,证毕。

由定理 4 知道区分集合满足最小覆盖的前提条件,即最小覆盖一定存在。

**定理 5**　$U_1$ 的一个集合覆盖等价于对应不完备信息系统的一个属性约简。

证明:　由区分集合的定义可以知道,任何一个 $U_1$ 的覆盖都可以把 $U_1$ 中的所有元素区别开来,即可以把对应不完备信息系统的所有相容类区别开来。根据不完备信息系统属性约简的定义可知,这些区分集合所对应的属性集就是对应不完备信息系统的一个属性约简。反之亦然。

由定理 5 我们可以知道如果 $U_1$ 的一个集合覆盖是最小覆盖,则其等价于不完备信息系统的一个最小约简。

**推论 1**　集合 $U_1$ 的最小集合覆盖对应于不完备信息系统的一个最小约简。

## 4　约简算法

任给不完备信息系统 $S = (U, AT, V, f)$

输入:不完备信息系统

输出:不完备信息系统最小属性约简

1) 求出 $U/SIM(AT)$;

2) 由定义构造关联矩阵 $\boldsymbol{M}$;

3) 由关联矩阵 $\boldsymbol{M}$ 求出所有属性的区分集合;

4) 把关联矩阵中的每个行向量看作一个元素,构成集合 $U$,求出集合 $U$ 的最小集合覆盖;

5) 输出 $U$ 的最小集合覆盖对应的属性集,即是最小约简。

算法分析:1) 由 $U/SIM(AT)$ 求相容类最坏情况下有 $|U|$ 个相容类,复杂度为 $O(|U|)$。2) 构造关联矩阵的复杂度为 $O(|U| * (|U|-1)/2) = O(|U|^2)$。3) 求区分集合时间复杂度为在 2) 的基础上考虑属性的个数 $|AT|$,复杂度

为 $O(|U|^2|AT|)$。4) 求最小集合覆盖复杂度在 3) 的基础上需要考虑最多 $|AT|$ 个属性的区分集合,复杂度为 $O(|U|^2|AT|^2)$,整个算法的时间复杂度为 $O(|U|^2|AT|^2)$。

## 5 实例分析

如表 1 的不完备信息系统,$U/SIM(AT) = \{X1, X2, X3, X4, X5, X6\}$,其中 $X1 = \{1\}$,$X2 = \{2,6\}$,$X3 = \{3\}$,$X4 = \{4,5\}$,$X5 = \{4,5,6\}$,$X6 = \{2,5,6\}$。差别矩阵如表 2。

**表 1 不完备信息系统示例**

| *Car* | *P* | *M* | *S* | *Ma* |
|---|---|---|---|---|
| 1 | High | High | Full | Low |
| 2 | Low | * | Full | Low |
| 3 | * | * | Compact | High |
| 4 | High | * | Full | High |
| 5 | * | * | Full | Hige |
| 6 | Low | High | Full | * |

**表 2 差别矩阵**

| $U_1$ | *P* | *M* | *S* | *Ma* | $U_1$ | *P* | *M* | *S* | *Ma* |
|---|---|---|---|---|---|---|---|---|---|
| $U1 = \{X1, X2\}$ | 1 | 0 | 0 | 0 | $U9 = \{X2, X6\}$ | 0 | 0 | 0 | 1 |
| $U2 = \{X1, X3\}$ | 0 | 0 | 1 | 1 | $U10 = \{X3, X4\}$ | 0 | 0 | 1 | 0 |
| $U3 = \{X1, X4\}$ | 0 | 0 | 0 | 1 | $U11 = \{X3, X5\}$ | 0 | 0 | 1 | 0 |
| $U4 = \{X1, X5\}$ | 1 | 0 | 0 | 1 | $U12 = \{X3, X6\}$ | 0 | 0 | 1 | 1 |
| $U5 = \{X1, X6\}$ | 1 | 0 | 0 | 1 | $U13 = \{X4, X5\}$ | 1 | 0 | 0 | 0 |
| $U6 = \{X2, X3\}$ | 0 | 0 | 1 | 1 | $U14 = \{X4, X6\}$ | 1 | 0 | 0 | 1 |
| $U7 = \{X2, X4\}$ | 1 | 0 | 0 | 1 | $U15 = \{X5, X6\}$ | 1 | 0 | 0 | 1 |
| $U8 = \{X2, X5\}$ | 1 | 0 | 0 | 1 | | | | | |

由相关矩阵我们可以看到属性 *M* 这一列全为零,这说明属性 *M* 对保持 $U/SIM(AT)$ 不变没有任何贡献,可以直接约去。

各属性的区分集合分别为:$S_P = \{U1, U4, U5, U7, U8, U13, U14, U15\}$;$S_M = \varnothing$;$S_S = \{U2, U6, U10, U11, U12\}$;$S_{Ma} = \{U2, U3, U4, U5, U6, U7, U8, U9, U12, U14, U15\}$。

由于集合 $S_P$, $S_S$, $S_{Ma}$ 三个集合是集合 $U_1$ 的最小集合覆盖,根据算法则对应的三个属性组成的集合 $\{P, S, Ma\}$ 就是原不完备信息系统的最小属性约简。

**参考文献:**


[1] 张文修, 吴伟志, 梁吉业, 等. 粗糙集理论与方法[M]. 北京: 科学出版社, 2001.1 - 40.

[2] WONG SKM, ZIARKO W. On optimal decision rules in decision tables[J]. Bulletin of Polish Academy of Sciences, 1985, 33(11/12): 693 - 696.

[3] HU XH, CERCONE N. Learning in relational database: A rough set approach[J]. International Journal of Computational Intelligence, 1995, 11(2): 323 - 338.

[4] JELONEK J, KRAWIEC K, SLOWINSKI R. Rough set reduction of attributes and their domains for neural networks[J]. International Journal of Computational Intelligence, 1995, 11(2): 339 - 347.

[5] GUAN JW, BELL DA. Rough computational methods for information systems[J]. Artificial Intelligences, 1998, 105(1/2): 77 - 103.

[6] 苗夺谦, 胡桂荣. 知识约简的一种启发式算[J]. 计算研究与发展, 1999, 36(6): 681 - 684.

[7] 王珏, 王任, 苗夺谦, 等. 基于 Rough Set 理论的"数据浓缩"[J]. 计算机学报, 1998, 21(5): 393 - 399.

[8] 周献中, 黄兵. 基于粗糙集的不完备信息系统属性约简[J]. 南京理工大学学报, 2003, 27(5): 630 - 635.

[9] 黄兵, 周献中, 张蓉蓉. 基于信息量的不完备信息系统属性约简[J]. 系统工程理论与实践, 2005, 25(4): 55 - 60.


(上接第 2663 页)

通过仿真性能测试与 augment UDDI Registry[2] 做了一些比较。试验结果表明,本系统的查准率(91%)略逊于 augment UDDI Registry 系统(97%),在查全率方面本系统(86%)明显高于 augment UDDI Registry(68%)。从分析得知,augment UDDI Registry 系统由于采用了基于 Prolog 推理的匹配算法,返回的结果是经过精确匹配的服务,所以具有很高的查准率。但是由于算法本身的匹配范围窄的局限,造成了查全率不高。本系统采用了"两阶段"匹配算法,综合考虑了服务的功能和服务质量的相似性。第一阶段的服务分类匹配剔除了与请求服务分类无关的广告服务,明显降低整个算法的时间复杂度,减少了第二阶段的匹配服务的个数,节省匹配时间,提高匹配引擎的效率,并且不会对整个算法的查全率和查准率造成影响,具有较好的匹配效果。

## 4 结语

本文通过引入服务质量本体,对现有的 Web 服务描述语言 OWL-S 进行了扩展,为基于语义相似度匹配算法提供了有效的服务描述语言。提出了"两阶段"匹配算法,该算法应用在原型系统 SWSC 上,试验结果表明较 augment UDDI Registry 系统的匹配方法能更大范围的定位服务,并且在效率方面有所提高,改善了 Web 服务发现的性能。

**参考文献:**


[1] TSALGATIDOU A, PILIOURA T. An Overview of Standards and Related Technology in Web Service[J]. Distributed and Parallel databases, 2002, 12(2/3): 135 - 162.

[2] PAOLUCCI M, KAWAMURA T, PAYNE TR, *et al*. Importing the semantic Web in UDDI[A]. Proceedings of Web Services, E-business and Semantic Web Workshop (CAiSE Workshop)[C]. Toronto, Canada, 2002. 225 - 236.

[3] PAOLUCCI M, KAWAMURA T, PAYNE TR, *et al*. Semantic Matching of Web Services Capabilities[A]. Proceedings of the 1st International Semantic Web Conference (ISWC)[C]. Sardinia, 2002. 333 - 347.

[4] PATIL A, OUNDHAKAR S, SHETH A, *et al*. METEOR-S Web service Annotation Framework[A]. The Proceedings of the Thirteenth International World Wide Web Conference[C], 2004. 553 - 562.

[5] 史忠植, 蒋运承, 张海俊, 等. 基于描述逻辑的主体服务匹配[J]. 计算机学报, 2004, 27(5): 625 - 635.

[6] 胡建强, 邹鹏, 王怀民, 等. Web 服务描述语言 QWSDL 和服务匹配模型研究[J]. 计算机学报, 2005, 28(4): 505 - 513.

[7] The OWL Services Coalition. Semantic markup for Web services (OWL-S)[S], 2004.

[8] ANDREA RM, EGENHOFER MJ. Determining semantic similarity among entity classes from different ontologies[J]. IEEE Transactions on Knowledge and Data Engineering, 2003, 15 (2): 442 - 456.

[9] http://projects.semwebcentral.org/frs/download.php/255/owls-tc2.zip[EB/OL], 2006.