不相容决策系统的知识库构造研究

黄治国[1],吴海涛[1],王加阳[2]
HUANG Zhi-guo[1],WU Hai-tao[1],WANG Jia-yang[2]

1.黄淮学院,河南 驻马店 463000
2.中南大学 信息科学与工程学院,长沙 410083
1.Huanghuai University,Zhumadian,Henan 463000,China
2.School of Information Science and Engineering,Central South University,Changsha 410083,China
E-mail:huangzhiguo2001@tom.com

HUANG Zhi-guo,WU Hai-tao,WANG Jia-yang.Study on construction of knowledge database for inconsistent decision system.Computer Engineering and Applications,2008,44(22):155-158.

Abstract: Rough set provides a formal theory model for construction of knowledge database,but it is worth studing in detail to construct knowledge database for inconsistent decision system.This paper defines distribution core and distribution reduction of a rule by applying notion of decision system's distribution reduction,and puts forward a kind of method based on distribution reduction for constructing knowledge database.This method gets the core of each condition class first,then achieves their distribution reductions by applying heuristic algorithm,and mines concise production rules for each condition class,constructs knowledge database for decision system.Furthermore,this method analyzes various situation that new objects are appended to the decision system,then updates current knowledge database incrementally,instead of running the whole constructins process again. This method would be applicable to inconsistent and consistent decision system.

Key words: rough set;construction of knowledge database;incremental updating

摘 要:粗糙集理论为知识库构造提供了一种形式化的理论模型,但是针对不相容决策系统构造知识库仍然是值得深入研究的问题。基于决策系统分布约简定义规则的分布核与分布约简概念,提出一种基于分布约简构造知识库的方法。首先确定各条件类的分布核,进而采用启发式算法计算其分布约简,挖掘约简规则集,构造出决策系统的知识库。并对加入决策系统中新对象的各种情形进行分析,对原有知识库进行增量式更新,而无需为更新知识库重新运行知识库构造算法。该方法能适应不相容决策系统,同样也适用于相容决策系统。

关键词:粗糙集;知识库构造;增量式更新
DOI:10.3778/j.issn.1002-8331.2008.22.046 **文章编号**:1002-8331(2008)22-0155-04 **文献标识码**:A **中图分类号**:TP274

1 引言

知识库与粗糙集理论属性约简具形式上的一致性,针对某特定的决策系统,一个约简可建立一个规则集,对应一个知识库。因此,粗糙集理论为知识库生成提供了一种形式化的理论模型。但是,针对不相容决策系统构造知识库仍然是值得深入研究的问题。

目前,已有较多学者就知识库构造方法进行了一定的研究。文献[2]提出一种最简知识库构造方法,可挖掘出满足给定精确度的最简产生式规则,该方法简洁有效,但是它通过人为地给定一分类正确度∂修改不一致对象决策值,最后所得知识库有可能与原决策系统不一致,并且也没有考虑知识库的增量式更新问题。文献[3]提出一种决策表约简的增量式学习方法,利用该方法得到分类规则知识库,但并没有考虑决策系统不相容的情形。文献[4]在等价矩阵概念基础上,提出进行数据清洗、提取决策规则的矩阵算法,但该方法通过数据清洗删除不一致对象会导致原决策系统信息的损失,最后所得知识库与原决策系统可能不一致。文献[5]讨论了最大分布约简、分配约简、分布约简和近似约简之间的关系,并给出了相应的可辨识矩阵,不协调目标决策系统的知识约简新方法,但并没有进一步给出知识库构造的具体方法。

本文基于决策系统分布约简定义规则的分布核与分布约简概念,提出一种基于分布约简构造知识库的方法。首先确定各条件类的分布核,进而采用启发式算法计算其分布约简,挖掘约简规则集,构造出决策系统的知识库。并对加入决策系统

基金项目:湖南省自然科学基金(the Natural Science Foundation of Hunan Province of China under Grant No.06JJ20075);河南省科技公关计划(the Key Technologies R&D Program of Henan Province,China under Grant No.0624220043)。

作者简介:黄治国(1978-),男,硕士,CCF 会员,主要研究方向:数据挖掘、决策支持、粗糙集理论及应用;吴海涛(1974-),男,讲师,主要研究方向:智能信息处理;王加阳(1963-),男,教授,博士,主要研究方向:智能计算、决策支持。

收稿日期:2007-10-09 **修回日期**:2008-03-03

中新对象的各种情形进行分析，对原有知识库进行增量式更新，而无需为更新知识库重新运行知识库构造算法。该方法能适应不相容决策系统，同样也适用于相容决策系统。

2 粗糙集基本概念

定义 1 一个目标决策信息系统(简称决策系统或信息系统)定义为四元组 $IS=<U,A,V,f>$。其中，$U=\{x_1,x_2,x_3,\cdots,x_n\}$为论域；$A=C\cup D$ 为有限属性集，C 和 D 分别为条件属性集和决策属性集；$V=\bigcup_{a\in A}V_a$，V_a 为属性 a 的值域；$f:U\times A\to V$ 为一信息函数使得 $f(x_i,a)\in V_A(x_i\in U,a\in A)$。

定义 2 属性集 $R(R\subseteq A)$上的不可分辨关系定义为：$IND(R)=\{(x_i,x_j)\in U^2:\forall a\in R,f(x_i,a)=f(x_j,a)\}$。论域 U 在属性集 C 上形成的划分 $U/IND(C)=\{X_1,X_2,\cdots,X_N\}$称为条件分类集，论域 U 在属性集 D 上形成的划分 $U/IND(D)=\{Y_1,Y_2,\cdots,Y_M\}$称为决策分类集。

$X\subseteq U$ 基于不可分辨关系 $IND(R)$的下近似定义为 $\underline{R}(X)=\cup\{X_i:X_i\in U/IND(R)\wedge X_i\subseteq X\}$，$\underline{R}(X)$是那些根据知识 R 判断肯定属于 X 的 U 中元素组成的集合；X 基于不可分辨关系 $IND(R)$的上近似定义为 $\overline{R}(X)=\cup\{X_i:X_i\in U/IND(R)\wedge X_i\cap X\neq\varnothing\}$，$\overline{R}(X)$是那些根据知识 R 判断可能属于 X 的 U 中元素组成的集合。

定义 3 决策系统 $IS=<U,A,V,f>$中，对于 $\forall x\in U$，对象 x 的决策规则定义为 $d_x:des([x]_{IND(C)})\to des([x]_{IND(D)})|(|[x]_{IND(C)}|,|[x]_{IND(C)}\cap[x]_{IND(D)}|)$。由于 $|[x]_{IND(C)}\cap[x]_{IND(D)}|=|[x]_{IND(C\cup D)}|$，因此对象 x 的决策规则又可表示为：$d_x:des([x]_{IND(C)})\to des([x]_{IND(D)})|(|[x]_{IND(C)}|,|[x]_{IND(C\cup D)}|)$。其中 $[x]_{IND(C)}\in U/IND(C)$为 x 关于条件属性集 C 上的等价类，$[x]_{IND(D)}\in U/IND(D)$为 x 关于决策属性集 D 上的等价类，$des([x]_{IND(C)})$和 $des([x]_{IND(D)})$分别表示对象 x 在不可分辨关系 $IND(C)$和 $IND(D)$上的描述。$|[x]_{IND(C)}|$与 $|[x]_{IND(C\cup D)}|$为修饰规则的两个参数，分别表示等价类 $[x]_{IND(C)}$的样本数目以及 $[x]_{IND(C\cup D)}$的样本数目。决策规则 d_x 关于 C 的约束 $d_x(C)$以及关于 D 的约束 $d_x(D)$，分别称为 d_x 的条件和决策。

决策系统中，若对于 $\forall x\in U$ 均有 $(|[x]_{IND(C)}|=|[x]_{IND(C\cup D)}|)$成立，则称规则 $d_x:des([x]_{IND(C)})\to des([x]_{IND(D)})|(|[x]_{IND(C)}|,|[x]_{IND(C\cup D)}|)$是相容的，否则称为是不相容的。若决策系统中所有规则均相容，则称此系统为相容决策系统，否则为不相容决策系统。

定义 4 决策系统 $IS=<U,A=C\cup D,V,f>$，$U=\{x_1,x_2,x_3,\cdots,x_n\}$，$U/IND(D)=(Y_1,Y_2,\cdots,Y_M)$，$B\subseteq C$，$x\in U$，称 $\mu_B(x)=(|Y_1\cap[x]_B|/|[x]_B|,|Y_2\cap[x]_B|/|[x]_B|,\cdots,|Y_M\cap[x]_B|/|[x]_B|)$为 x 关于 B 的分布函数。显然，$\mu_B(x)$是 $[x]_B$ 中元素在 $U/IND(D)$上的概率分布。若 $(\forall x\in U)(\mu_B(x)=\mu_C(x))$，则称 B 是分布协调集。若 B 是分布协调集且 B 的任意真子集都不是分布协调集，则称 B 为分布约简。

3 基于分布约简的知识库构造

决策系统核属性的确定对属性约简具重要意义，一直受到粗糙集理论界学者的关注。Hu 在文献[6]中根据 Skowron 提出的可辨识矩阵得出一个确定决策系统核属性集的方法。叶东毅教授在文献[7]中对 Hu 的结论提出质疑，并通过改进可辨识矩阵提出了一种计算核属性的方法。王国胤教授在文献[8]中对上述两种方法进行分析，分别指出其局限性，并提出一种决策表信息熵定义下的核属性计算方法。文献[8]指出：对于相容决策系统可采用 Hu 的方法计算核属性，对于不相容决策系统可采用叶的方法计算核属性，而无论决策系统是否相容均可使用信息熵定义下的核属性计算方法。

本章将基于文献[5]所提出分布约简概念定义决策表的分布核属性，并与文献[8]中信息熵定义下的核属性进行比较研究。然后定义条件类的分布核与分布约简，阐述基于分布约简构造知识库的基本原理。

定义 5 决策系统 $IS=<U,A=C\cup D,V,f>$，$U=\{x_1,x_2,x_3,\cdots,x_n\}$，$a\in C$，决策系统 IS 的分布核属性集表示为 $CORE_D(C)$，a 为分布核属性即 $a\in CORE_D(C)$当且仅当 $(\exists x\in U)(\mu_{C-\{a\}}(x)\neq\mu_C(x))$。

分布核属性集 $CORE_D(C)$是保持所有对象关于每个决策类的隶属程度不变的属性集的必要组成部分，删除核属性集的任意子集必导致部分对象关于某些决策类的隶属程度发生改变。

定理 1 $CORE_D(C)=\cap RED_D(C)$，其中 $RED_D(C)$是决策表的所有分布约简族。

证明

(1)若 $a\in CORE_D(C)$，则 $(\exists x\in U)(\mu_{C-\{a\}}(x)\neq\mu_C(x))$；设 $B\subseteq C$ 为 C 关于 D 的任意一个分布约简，则 $(\forall x\in U)(\mu_B(x)=\mu_C(x))$。因而有 $(\exists x\in U)(\mu_{C-\{a\}}(x)\neq(\mu_B(x))$，即 $B\subseteq C-\{a\}$，又由于 $B\subseteq C$，所以 $a\in B$。因此若 $a\in CORE_D(C)$则必有 $a\in\cap RED_D(C)$，即 $CORE_D(C)\subseteq\cap RED_D(C)$。

(2)若 $a\in\cap RED_D(C)$，运用反证法假设 $a\notin CORE_D(C)$，则据定义 5 有 $(\forall x\in U)(\mu_{C-\{a\}}(x)=\mu_C(x))$，那么至少存在 C 关于 D 的一个分布约简 B 且 $B\subseteq C-\{a\}$，所以至少存在 C 关于 D 的一个分布约简 B 且 $a\notin B$，这与前提条件 $a\in\cap RED_D(C)$矛盾，因此若 $a\in\cap RED_D(C)$，则必有 $a\in CORE_D(C)$。即 $\cap RED_D(C)\subseteq CORE_D(C)$。

证毕。

定理 1 表明，分布核概念的意义主要体现在两个方面：首先分布核可解释为决策表中不能消去的属性集，因为缺少核属性将导致部分对象关于某些决策类的隶属程度发生改变；其次分布核可以作为分布约简的计算基础，因为分布核包含在所有的分布约简之中，并且计算可以直接进行。

定义 6 知识(属性集合)P 的熵 $H(P)$定义为：$H(P)=\sum_{i=1}^{|U/IND(P)|}(|U/IND(P)[i]|/|U|)\log(|U/IND(P)[i]|/|U|)$。其中 $U/IND(P)$为论域 U 在不可分辨关系 $IND(P)$导致的划分，$U/IND(P)[i]$为此划分中第 i 个等价类。

定义 7 知识(属性集合)$Q(U/IND(Q)=\{Y_1,Y_2,\cdots,Y_m\})$相对于知识(属性集合)$P(U/IND(P)=\{X_1,X_2,\cdots,X_n\})$的条件熵 $H(Q|P)$定义为：$H(Q|P)=-\sum_{i=1}^{n}P(X_i)\sum_{j=1}^{m}P(Y_j|X_i)\log(P(Y_j|X_i))$，其中，$P(X_i)=|X_i|/U$，$P(Y_j|X_i)=|X_i\cap Y_j|/|X_i|$，$i=1,2,\cdots,n$，$j=1,2,\cdots,m$。

定义 8 给定一决策信息系统 $IS=<U,A=C\cup D,V,f>$，$B\subset C$，则任意属性 $a\in C-B$ 的重要性定义为：$SGF(a,B,D)=H(D|B)-H(D|B\cup\{a\})$。

若 $B=\varnothing$，则 $SGF(a,B,D)=H(D)-H(D|\{a\})$称为属性 a 和决策 D 的互信息，记为 $I(a,D)$。$SGF(a,B,D)$的值越大，说明在已知 B 的条件下，a 对于决策 D 就越重要。

定义 9 设 $IS=<U,A=C\cup D,V,f>$为一决策系统，则 $a\in C$ 是

C 相对于 D 在信息熵定义下的核属性当且仅当 $H(D|C)\neq H(D|C-\{a\})$。

定理 2 设 $IS=<U,A=C\cup D,V,f>$为一决策系统，则对于 $\forall a\in C$ 有:$(\exists x\in U)(\mu_{C-\{a\}}(x)\neq\mu_C(x))\Leftrightarrow H(D|C)\neq H(D|C-\{a\})$。

为证明上述定理,还需引用文献[9]中的引理 1。

引理 1 设论域为 U,U 在某个等价关系上形成的划分为 $A_1=\{X_1,X_2,\cdots,X_n\}$,而 $A_2=\{X_1,\cdots,X_{i-1},X_{i+1},\cdots,X_{j-1},X_{j+1},\cdots,X_n,X_i\cup X_j\}$ 是将划分 A_1 中的某两个等价块 X_i 与 X_j 合并得到的新划分。$B=\{Y_1,Y_2,\cdots,Y_m\}$也是 U 上的一个划分。且记 $H(B|A_1)=-\sum_{i=1}^{n}P(X_i)\sum_{j=1}^{m}P(Y_j|X_i)\log(P(Y_j|X_i))$,$H(B|A_2)=H(B|A_1)-P(X_i\cup Y_j)\sum_{k=1}^{m}P(Y_k|X_i\cup Y_j)\log(P(Y_k|X_i\cup Y_j))+P(X_i)\sum_{k=1}^{m}P(Y_k|X_i)\log(P(Y_k|X_i))+P(X_j)\sum_{k=1}^{m}P(Y_k|X_j)\log(P(Y_k|X_j))$,则 $H(B|A_2)\geqslant H(B|A_1)$。

引理 1 表明，如果将决策表条件属性的分类进行合并,那么将导致条件熵的单调上升,只有在发生合并的两个分类对于决策类的隶属度(概率)相等的情况下,才不会导致条件熵的变化。其次,划分 $U/IND(C-\{a\})$是可以通过将划分 $U/IND(C)$中的部分等价块合并得到的,如果 $H(D|C-\{a\})=H(D|C)$,则所有被合并在一起的等价块对于决策类的隶属度均相等,因此,在合并后,条件属性分类中的等价块对于各个决策属性分类的隶属度不会发生变化。

下面证明定理 2。

证明 必要性。若$(\exists x\in U)(\mu_{C-\{a\}}(x)\neq\mu_C(x))$,则说明删除属性 a 引起划分 $U/IND(C)$中某些等价块发生了合并,且存在被合并的等价块对于决策类的隶属度不同的情形,则据引理 1 知条件熵必定发生变化,即 $H(D|C)\neq H(D|C-\{a\})$。因此$(\exists x\in U)(\mu_{C-\{a\}}(x)\neq\mu_C(x))\Rightarrow H(D|C)\neq H(D|C-\{a\})$。

充分性。证其逆否命题$(\forall x\in U)(\mu_{C-\{a\}}(x)=\mu_C(x))\Rightarrow H(D|C)=H(D|C-\{a\})$。若$(\forall x\in U)(\mu_{C-\{a\}}(x)=\mu_C(x))$,则只可能存在两种情形:(1)条件等价类无合并情形;(2)条件等价类有合并情形,但这些合并的条件等价类相对于决策类的隶属度不变。对于情形(1)由定义 7 可知其条件熵不会发生变化,对于情形(2),由引理 1 可知其条件熵也不会发生变化。因此命题$(\forall x\in U)(\mu_{C-\{a\}}(x)=\mu_C(x))\Rightarrow H(D|C)=H(D|C-\{a\})$成立,故其逆否命题同样成立即 $H(D|C)\neq H(D|C-\{a\})\Rightarrow(\exists x\in U)(\mu_{C-\{a\}}(x)\neq\mu_C(x))$。充分性得证。证毕。

定理 2 说明基于分布约简定义的核属性与信息熵理论下定义的核属性均是保证决策表条件等价类中对象在决策类集中的概率不发生变化,它们是等价的。

定义 10 条件类 $X_i\subseteq U$ 的分布核定义为:$CORE_D(C)[i]=\{a:(\forall x\in X_i)\wedge(\mu_{C-\{a\}}(x)\neq\mu_C(x))\}$。即条件类分布核 $CORE_D(C)[i]$中属性对条件类 X_i 而言是不可缺少的，否则会导致 X_i 中对象关于某些决策类的隶属程度发生改变。

定义 11 条件类 $X_i\subseteq U$ 的分布协调集定义为:$\{B:((\forall x\in X_i)\wedge(B\subseteq C)\wedge(\mu_C(x)=\mu_B(x)))\wedge((\forall X_j\forall y)(i\neq j)\wedge(y\in X_j)\wedge(des([y]_B)=(des([x]_B)\rightarrow\mu_C(y)=\mu_B(y)))\}$。若 B 是条件类 $X_i\subseteq U$ 的分布协调集且 B 的任意真子集都不是 X_i 的分布协调集,则称 B 为 X_i 的分布约简,记为 $RED_D(C)[i]$。

这说明若 B 是条件类 X_i 的分布约简,则 B 与 C 针对X_i 中对象关于每一决策类具有同样的隶属程度。可用 B 代替 C 来产生 X_K 的分类规则,所得规则即为简化的规则。

生成条件类 X_i 的分类规则的原则为：针对所有决策类 $U/IND(D)=\{Y_1,Y_2,\cdots,Y_M\}$计算$|[X_i]_{IND(C)}\cap Y_j|$,若$|[X_i]_{IND(C)}\cap Y_j|\neq 0$ 则生成规则 $des([X_i]_{IND(C)})\rightarrow des([X_i]_{IND(D)})|(|[X_i]_{IND(C)}|,|[X_i]_{IND(C)}\cap Y_j|)$。若 B 是条件类 X_i 的分布约简,可将其化简后得到:$des([X_i]_{IND(B)})\rightarrow des([X_i]_{IND(D)})|(|[X_i]_{IND(B)}|,|[X_i]_{IND(B)}\cap Y_j|)$。

定理 3 条件类 X_i 的分类规则化简过程中,若存在其它条件类的分类规则与之合并,则规则的参数必发生变化即$(|[X_i]_{IND(C)}|,|[X_i]_{IND(C\cup D)}|)\neq(|[X_i]_{IND(B)}|,|[X_i]_{IND(B\cup D)}|)$,但其支持度必不发生变化,即$|[X_i]_{IND(C\cup D)}|/|[X_i]_{IND(C)}|=|[X_i]_{IND(B\cup D)}|/|[X_i]_{IND(B)}|$。

证明（运用反证法）条件类 X_i 的分类规则化简过程中,假设参数$|[X_i]_{IND(B)}|$与$|[X_i]_{IND(B\cup D)}|$不发生变化,则不会存在任意其它条件类与 X_i 发生合并的情形，这与前提条件相矛盾因此两个参数必发生变化。

条件类 X_i 的分类规则化简过程中,假设$|[X_i]_{IND(C\cup D)}|/|[X_i]_{IND(C)}|\neq|[X_i]_{IND(B\cup D)}|/|[X_i]_{IND(B)}|$,又$|[X_i]_{IND(C\cup D)}|/|[X_i]_{IND(C)}|$与$|[X_i]_{IND(B\cup D)}|/|[X_i]_{IND(B)}|$分别为分布函数 $\mu_C(x)$与 $\mu_B(x)$中相对应的一项$(x\in X_i,B\subseteq C)$,这表明 $\mu_C(x)\neq\mu_B(x)$;而 X_i 的分类规则化简过程只会在 B 为 X_i 的分布约简前提即 $\mu_C(x)=\mu_B(x)$条件下进行。$\mu_C(x)\neq\mu_B(x)$与 B 为 X_i 的分布约简前提相矛盾。因此规则的支持度必不发生变化即$|[X_i]_{IND(C\cup D)}|/|[X_i]_{IND(C)}|=|[X_i]_{IND(B\cup D)}|/|[X_i]_{IND(B)}|$。证毕。

定理 3 表明,在条件类的分类规则化简过程中,规则的支持度不会发生变化。这意味着,只有当不存在条件等价类合并,或所合并条件等价类的分布函数相等,两种情形必存在其中之一时才可进行化简,否则相应规则不能化简。

根据上述定理与定义,可设计基于分布约简的知识库构造算法如下。

算法 1 基于分布约简的知识库构造算法

输入:决策系统 $IS=<U,A=C\cup D,V,f>$

输出:知识库

步骤 1 计算 U 关于 C 的等价类 $U/IND(C)$,U 关于 D 的等价类 $U/IND(D)$;

步骤 2 计算每个条件类 $X_i\in U/IND(C)$的分布核属性集 $CORE_D(C)[i]$;

(1)for each $X_i\in U/IND(C)$ do

{$CORE_D(C)[i]=\emptyset$;

for each $a\in C$ do

if $((\forall x\in X_i)(\mu_{C-\{a\}}(x)\neq\mu_C(x))$

$CORE_D(C)[i]=CORE_D(C)[i]+a$;

}

步骤 3 针对每个条件类 X_i,计算其分布约简;

(1)for each $X_i\in U/IND(C)$ do

{$RED_D(C)[i]=CORE_D(C)[i]$;

$sel=C-RED_D(C)[i]$;

while $RED_D(C)[i]$不为 X_i 的分布约简)

{for each $a\in sel$ 计算其属性重要性 $SGF(a,RED_D(C)[i],D)$;

$a=SGF$ 值最大的属性;

$RED_D(C)[i]=RED_D(C)[i]+\{a\}$;

$sel=sel-\{a\}$

}

}

步骤 4 针对每个条件类 X_i,生成其分类规则;

(1)$KDB=\emptyset$;

(2)for each $X_i\in U/IND(C)$ do

{$B=RED_D(C)[i]$};

for each $Y_j \in U/IND(D)$ do

if($X_i \cap Y_j \neq \varnothing$) $KDB=RKDB+\{des([X_i]_{IND(B)}) \rightarrow des(Y_j)|(|[X_i]_{IND(B)}|, |[X_i]_{IND(B)} \cap Y_j|)\}$;

}

步骤 5 结束。

求条件类的最小分布约简是一个完全问题,算法 1 利用属性重要性作为启发函数求近似最小约简。其基本过程是,首先得到条件类的分布核作为求取其分布约简的基础,然后按照属性的重要程度从大到小逐个加入属性,直至得到其分布约简为止。无论决策系统是否相容,此知识库构造方法均适用,且所构造知识库保持与原决策系统一致。

4 知识库的增量式更新

当有新对象加入决策系统时,就需要对出现的各种新对象进行分析,更新现有的知识库。假设决策系统 IS 在条件属性集 C 与决策属性集 D 上形成的划分分别为 $U/IND(C)=\{X_1, X_2, \cdots, X_n\}$和 $U/IND(D)=\{Y_1, Y_2, \cdots, Y_m\}$。新对象 x 加入后原知识库中知识的形式和参数均有可能发生变化。此过程通过区分以下几种情形来实现:

(1)$x \notin X_i(i=1,2,\cdots,n)$且 $y \notin Y_j(j=1,2,\cdots,m)$。

对象 x 形成了一新的条件类 X_{n+1} 和新的决策类 Y_{m+1},首先计算新条件类 X_{n+1} 的分布核, 然后根据属性重要度函数求得其分布约简 $RED_D(C)[n+1]$。在原知识库中加入新规则 $\{des([X_{n+1}]_{IND(RED_D(C)[n+1])}) \rightarrow des(Y_{m+1})|(|[X_{n+1}]_{IND(RED_D(C)[n+1])}|, |[X_{n+1}]_{IND(RED_D(C)[n+1])} \cap Y_{m+1}|)\}$。因为对象 x 形成了一新的条件类 X_{n+1},所以此规则的两个参数均为 1。

(2)$x \notin X_i(i=1,2,\cdots,n)$且 $y \in Y_j(j=1,2,\cdots,m)$。

对象 x 形成了一新的条件类 X_{n+1}, 首先计算新条件类 X_{n+1} 的分布核, 然后根据属性重要度函数求得其分布约简 $RED_D(C)[n+1]$。在原知识库中加入新规则$\{des([X_{n+1}]_{IND(RED_D(C)[n+1])}) \rightarrow des(Y_{m+1})|(|[X_{n+1}]_{IND(RED_D(C)[n+1])}|, |[X_{n+1}]_{IND(RED_D(C)[n+1])} \cap Y_{m+1}|)\}$。因为此情形下对象 x 形成了一新的条件类 X_{n+1},所以此规则的两个参数也均为 1。

(3)$x \in X_i(i=1,2,\cdots,n)$且 $y \notin Y_j(j=1,2,\cdots,m)$。

对象 x 形成了一新的决策类 Y_{m+1} 但没有形成新的条件类,算法 1 已得到此条件类的分布约简 $RED_D(C)[i]$。针对原知识库中所有前件为 $des([X_i]_{IND(RED_D(C)[i])})$的规则,更新其参数令$|[X_i]_{IND(RED_D(C)[i])}|=|[X_i]_{IND(RED_D(C)[i])}|+1$;并加入一条新规则 $\{des([X_i]_{IND(RED_D(C)[i])}) \rightarrow des(Y_{m+1})|(|[X_i]_{IND(RED_D(C)[i])}|, 1)\}$。

(4))$x \in X_i(i=1,2,\cdots,n)$且 $y \in Y_j(j=1,2,\cdots,m)$。

对象 x 没有形成新的决策类 Y_{m+1} 也没有形成新的条件类,算法 1 已得到此条件类的分布约简 $RED_D(C)[i]$。针对原知识库中所有前件为 $des([X_i]_{IND(RED_D(C)[i])})$的规则,更新其参数令$|[X_i]_{IND(RED_D(C)[i])}|=|[X_i]_{IND(RED_D(C)[i])}|+1$;针对原知识库中前件为 $des([X_i]_{IND(RED_D(C)[i])})$后件为 $des(Y_j)$的规则,更新其参数令$|[X_i]_{IND(RED_D(C)[i])} \cap Y_j|=|[X_i]_{IND(RED_D(C)[i])} \cap Y_j|+1$。

新对象 x 加入后,通过匹配以上情形即可实现知识库的增量式更新。

5 结论

本文基于决策系统分布约简定义规则的分布核与分布约简概念,提出一种基于分布约简的构造知识库的方法,它能适应决策系统的不一致性,首先确定各条件类的分布核,进而采用启发式算法计算其分布约简,挖掘约简规则集,构造出决策系统的知识库。并对加入决策系统中新对象的各种情形进行分析,对原有知识库进行增量式更新,而无需为更新知识库重新运行知识库构造算法。该方法的优点是能适应决策系统的不相容情形,且运用此方法所获取知识库能够保持与原决策系统一致;同时,针对加入决策系统中新对象的各种情形在原知识库基础上进行增量式更新,避免了为更新知识库而重新知识库构造算法。

参考文献:

[1] Pawlak Z,Grzymala-Busse J,Slowinski R,et al.Rough sets[J].Communications of the ACM,1995,38(11):89-95.

[2] 赛煜,王海洋.一种基于粗糙集理论的最简规则挖掘方法[J].计算机工程,2003,29(20):77-79.

[3] 李滔,王俊普,徐杨.一种基于粗糙集的网页分类方法[J].小型微型计算机系统,2003,24(3):520-522.

[4] 谭天乐,宋执环,李平.信息系统数据清洗、规则提取的矩阵算法[J].信息与控制,2003,32(4):289-294.

[5] 张文修,米据生,吴伟志.不协调目标信息系统的知识约简[J].计算机学报,2003,26(1):12-18.

[6] Hu X H,Cercone N.Learning in relational databases:a rough set approach[J].Computational Intelligence,1995,11(2):323-337.

[7] 叶东毅,陈昭炯.一个新的差别矩阵及其求核方法[J].电子学报,2002,30(7):1086-1088.

[8] 王国胤.决策表核属性的计算方法[J].计算机学报,2003,26(5):611-615.

[9] 王国胤,于洪,杨大春.基于条件信息熵的决策表约简[J].计算机学报,2002,25(7):759-766.

(上接 154 页)

检索时的满意度。

虽然本文从语义关系角度对多关键词检索作了有益的探索,但是仍存在不足之处,有待于进一步探索研究:其中,考虑到检索速度的要求,采用了简化的多关键词关系与核心关键词提取算法,使得一些结果不太令人满意。

参考文献:

[1] 董振东,董强.知网[EB/OL].http://www.keenage.corn.

[2] 董振东.语义关系的表达和知识系统的建造[J].语言文字应用,1998,27(3):76-82.

[3] 陈伟雄,马少平.基于元搜索引擎的多关键词检索技术[J].计算机工程与应用,2004,40(24):83-84.

[4] 龚永恩,袁春风.基于语义的词义消歧算法初探[J].计算机应用研究,2006,23(3):42-43.

[5] 许云,樊孝忠.基于知网的语义相关度计算[J].北京理工大学学报,2005,25(5):412-413.

[6] 李素建.基于语义计算的语句相关度研究[J].计算机工程与应用,2002,38(7):75-76.

[7] 刘群,李素建.基于《知网》的词汇语义相似度计算[J].计算语言学及中文信息处理,2002,7:59-76.

[8] 夏天.汉语词语语义相似度计算研究[J].计算机工程,2007,33(6):191-192.