概念的属性约简及异构数据概念漂移探测

邓大勇[1,3],卢克文[1],黄厚宽[2],邓志轩[1]
(1. 浙江师范大学数理与信息工程学院 浙江金华 321004;2. 北京交通大学计算机与信息技术学院 北京 100044;
3. 浙江师范大学行知学院 浙江金华 321004)

摘 要: 粗糙集是粒计算的一种重要方法,数据异构性是大数据的一种特征.针对异构数据问题,探索了粗糙集属性约简的本质,提出了概念属性约简的定义,它兼容值约简、Pawlak约简和并行约简.探究了概念属性约简的性质,提出了异构数据的属性约简方法和概念漂移探测方法.理论分析和示例表明了这些方法的有效性.为粗糙集、粒计算融入大数据的时代潮流提供了一种新方法.

关键词: 粒计算;F-粗糙集;属性约简;异构数据;概念漂移
中图分类号: TP18 **文献标识码:** A **文章编号:** 0372-2112 (2018)05-1234-06
电子学报URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2018.05.032

Attribute Reduction for Concepts and Concept Drifting Detection in Heterogeneous Data

DENG Da-yong[1,3], LU Ke-wen[1], HUANG Hou-kuan[2], DENG Zhi-xuan[1]
(1. *College of Mathematics, Physics and Information Engineering, Zhejiang Normal University, Jinhua, Zhejiang* 321004, *China*;
2. *School of Computer and Information Technology, Beijing Jiaotong University, Beijing* 100044, *China*;
3. *Xingzhi College, Zhejiang Normal University, Jinhua, Zhejiang* 321004, *China*)

Abstract: Rough set theory is one of important methods of granular computing, and data heterogeneities are one of remarkable characteristics in big data. For data heterogeneities, we define attribute reduction for concepts after investigating intrinsic quality of attribute reducts, which can contain value reducts, Pawlak attribute reducts and parallel reducts. After investigating properties of concept-attribute-reduction, we present a new method to reduce redundant attributes and a new method to detect concept drift for heterogeneous concepts. Theoretical analysis and examples show that these methods are valid. This work provides a new type way for rough set theory and granular computing to integrate into big data.

Key words: granular computing; F-rough sets; attribute reduction; heterogeneous data; concept drifting

1 引言

粗糙集[1]是粒计算的重要方法.它最主要的应用在于属性约简和不确定性分析,最大的优势在于不需要先验知识.F-粗糙集[2]是第一个动态粗糙集模型,并行约简是与其相对应的属性约简方法.

在大数据时代,不少粒计算和粗糙集的研究者试图将粗糙集理论融入大数据的时代潮流,也取得了一些研究成果.MapReduce技术被用于大规模数据的并行计算和属性约简[3].粗糙集属性约简方法用于机器学习或智能算法前的数据预处理[4].基于桶排序策略的快速属性约简方法被用于邻域粗糙集模型中[5].将分而治之的策略与粗糙集方法相结合[6].粗糙集的属性约简方法用之于在线特征选择[7].用大规模粒计算方法统一粗糙集[8].更值得关注的是,文[9~11]把粗糙集理论与数据流挖掘相结合,用粗糙集方法探测概念漂移.文[12]对当前粒计算与大数据相结合的工作进行了总结,提出了3大挑战,特别是异构数据的处理.

运用粒计算、粗糙集方法处理异构数据的文献比较少见.文[13,14]针对混合数据类型数据提出了基于粗糙集、粒计算的处理方法,是属性集相同、数据类型不同的异构数据处理方法.多粒度粗糙集[15]用不同异构属性子集表示单个信息表中的不确定性,形式上是异构的,但基本上还没能用于异构数据的约简或数据挖

收稿日期:2016-12-16;修改日期:2017-03-21;责任编辑:蓝红杰
基金项目:国家自然科学基金(No.61473030);浙江省自然科学基金(No.LY15F020012);浙江师范大学网络空间安全浙江省一流学科

掘. 针对文[12]提出的问题,结合我们的思考:属性约简的主体是什么? 属性约简的本质是什么? 能否用粗糙集的方法处理异构数据(属性集不同)并探测概念漂移?

针对这几个问题,本文努力给出我们的回答. 传统的概念漂移[16~19]是由时间变化引发的,往往用概念漂移的结果(比如:分类准确率)来探测概念漂移. 文[11]扩展了概念漂移的定义,研究了由空间或条件变化引发的概念漂移. 文[10,11]探索了使用概念内部表示的变化(比如:属性重要性变化)来探测概念漂移. 继承文[11]的基本思想,本文从粗糙集最根本的"概念"出发,探究属性约简的本质. 定义了概念的属性约简,它兼容值约简、Pawlak 属性约简和并行约简. 并把概念属性约简方法运用于异构数据(属性集不同)的属性约简和概念漂移探测. 首先,定义了概念集的属性依赖度,并由此定义了概念集的属性重要性,这两个定义兼容 Pawlak 和并行约简的属性依赖度和属性重要性,并进行了推广. 其次,定义了概念集的属性约简,它兼容值约简、Pawlak 属性约简和并行约简;探究了概念集属性约简的性质. 再次,将概念集属性约简运用于异构数据(属性集不同)的属性约简. 最后,定义了异构数据的质概念漂移和量概念漂移,扩展了概念漂移的意义. 为粗糙集理论进一步融入大数据的时代潮流做了一定的理论奠基.

2　粗糙集与 F-粗糙集

假设读者比较熟悉粗糙集,本节仅简单介绍粗糙集[1]和 F-粗糙集[2]的定义.

设 $K=(U,A)$ 是一个知识系统,其中 U 为论域,A 为等价关系簇(或属性集合).

概念 $Y\subseteq U$ 相对于等价关系簇 A 的上、下近似定义为:

$$\overline{A}(K,Y)=\cup\{[x]_A:[x]_A\cap Y\neq\varnothing\},$$

$$\underline{A}(K,Y)=\cup\{[x]_A:[x]_A\subseteq Y\}.$$

其中$[x]_A$ 是等价类. $(\underline{A}(K,Y),\overline{A}(K,Y))$称为粗糙集.

设 $F=\{K_i:K_i=(U_i,A),i=1,2,\cdots,n\}$ 是一个知识系统簇,$Y(K_i)\subseteq U_i$ 是一个在不同的知识系统中表示可能不同的概念或称为一个在不同的知识系统 K_i 中意义有所变化的概念变量. 如果不引起混淆,$Y(K_i)$可以简记为 Y. 则 Y 在知识系统簇中的上、下近似定义为:

$$\overline{A}(F,Y)=\{\overline{A}(K,Y):K\in F\},$$

$$\underline{A}(F,Y)=\{\underline{A}(K,Y):K\in F\}.$$

$(\underline{A}(F,Y),\overline{A}(F,Y))$称为 F-粗糙集.

注:知识系统有时也被称为信息系统,如果出现决策属性,则称其为决策系统.

由此可见,粗糙集的基础是概念. 下文将从概念或概念簇的角度探究属性约简.

3　属性依赖度及属性重要性的归一化

定义 1　设 $S=\{Y_1,Y_2,\cdots,Y_N\}$ 是决策系统簇 $F=\{DS_i:DS_i=(U_i,A_i,d_i),i=1,2,\cdots,n\}$ 中的概念集,即 $Y_j\in\cup U_i/\{d_i\}(j=1,2,\cdots,N.\ i=1,2,\cdots,n)$. $A=\cup_{i=1}^{n}A_i$ 是概念集合 S 所涉及的全部条件属性集合,$D=\cup_{i=1}^{n}\{d_i\}$ 是概念集合 S 所涉及的全部决策属性集合,称决策属性集 D 以程度 $h(0\leqslant h\leqslant 1)$ 依赖条件属性集 A,其中,$h=\gamma(S,A,D)=\dfrac{\sum_{Y\in S}|\underline{A}(Y)|}{\sum_{Y\in S}|Y|}$.

定义 2　设 $S=\{Y_1,Y_2,\cdots,Y_N\}$ 是决策系统簇 $F=\{DS_i:DS_i=(U_i,A_i,d_i),i=1,2,\cdots,n\}$ 中的概念集,$A=\cup_{i=1}^{n}A_i$ 是概念集合 S 所涉及的全部条件属性集合,$D=\cup_{i=1}^{n}\{d_i\}$ 是概念集合 S 所涉及的全部决策属性集合. $B\subseteq A$ 为属性子集,对于任意的 $\alpha\in A$ 相对于 B 的属性重要性定义为:

(1)内属性重要性为 $\sigma(S,B,\alpha)=\gamma(S,B,D)-\gamma(S,B-\{\alpha\},D)$;

(2)外属性重要性为 $\sigma'(S,B,\alpha)=\gamma(S,B\cup\{\alpha\},D)-\gamma(S,B,D)$.

4　概念集的属性约简

由概念或概念簇出发,属性约简的定义可以更一般地推广.

定义 3　设 $S=\{Y_1,Y_2,\cdots,Y_N\}$ 是知识系统簇 $F=\{K_i:K_i=(U_i,A_i),i=1,2,\cdots,n\}$ 中的概念集,即 $Y_j\subseteq U_i$ $(j=1,2,\cdots,N.\ i=1,2,\cdots,n)$,$A=\cup_{i=1}^{n}A_i$ 是概念集合 S 所涉及的全部条件属性集合,则条件属性集 $B\subseteq A$ 是 S 的属性约简 iff B 满足以下两个条件:

(1)对于任意的 $Y\in S$ 都有$\underline{A}(Y)=\underline{B}(Y)$;

(2)对于任意的 $C\subset B$,都存在 $Y\in S$ 使得$\underline{A}(Y)\neq\underline{C}(Y)$;

S 的所有属性约简记为 $Red(S)$,$Core(S)=\cap Red(S)$称为 S 的属性核.

定义 3 不仅继承了值约简、Pawlak 约简与并行约简的基本思想,而且兼容这 3 个约简的定义. 当集合 S 中仅有一个关于决策属性的概念时,此定义为值约简;当 S 中的元素为一个决策属性的全部概念时,此定义为 Pawlak 约简;当 S 中的元素为一个决策表簇中全部决策属性的概念时,此定义为并行约简.

此外,这个约简定义扩展了约简的范围,不仅单个决策属性的概念可以约简,而且不同决策属性的多个概念也可以约简,甚至异构数据也能约简.

下面研究概念集属性约简的性质:

定理 1 对于概念集 $S=\{Y_1,Y_2,\cdots,Y_N\}$ 和 $Y\in S$ 来说,如果 $\alpha\in\cap Red(Y)$,则 $\alpha\in\cap Red(S)$;反过来,如果 $\alpha\in\cap Red(S)$,则存在 $Y\in S$ 使得 $\alpha\in\cap Red(Y)$.

定理 2 对于概念集 $S=\{Y_1,Y_2,\cdots Y_N\}$ 来说,$\alpha\in\cap Red(S)$ 当且仅当 α 相对于 A 的内属性重要性大于 0,即:$\sigma(S,A,\alpha)>0$.

推论 1 设概念集 $S=\{Y_1,Y_2,\cdots Y_N\}$,下面两个结论成立:

(1)对于任意的 $\alpha\in A$,内属性重要性 $\sigma(S,A,\alpha)=0$ 当且仅当对于任意 $Y_i\in S$ 有 $\sigma(Y_i,A,\alpha)=0$;

(2)对于任意的 $\alpha\in A$,内属性重要性 $\sigma(S,A,\alpha)>0$ 当且仅当存在 $Y_i\in S$ 使得 $\sigma(Y_i,A,\alpha)>0$.

定理 3 对于概念集 $S=\{Y_1,Y_2,\cdots Y_N\}$ 来说,$B\subseteq A$ 是 S 的属性约简,则对于任何的 $\alpha\in B$ 相对于 B 的内属性重要性不小于相对于 A 的内属性重要性,即:$\sigma(S,B,\alpha)\geqslant\sigma(S,A,\alpha)$.

定理 4 设概念集 $S=\{Y_1,Y_2,\cdots Y_N\}$,$B\subseteq A$ 是概念集 S 的属性约简,则对于任意的 $\alpha\in B$ 有 $\sigma(S,B,\alpha)>0$,对于任意的 $\alpha\in(A-B)$ 有 $\sigma'(S,B,\alpha)=0$.

定理 5 设概念集 $S=\{Y_1,Y_2,\cdots Y_N\}$,$B\subseteq A$. 对于任意的 $\alpha\in(A-B)$,如果 $\sigma'(S,B,\alpha)=0$,则条件属性 α 相对 B 是可约去的.

例 1 设 $F=\{DT_1,DT_2\}$ 是一个决策子表簇,如表 1、表 2 所示,a、b、c 是条件属性,d 是决策属性.

表 1 决策表 DT_1

U_1	a	b	c	d
x_1	0	0	1	1
x_2	1	1	0	1
x_3	0	1	0	0
x_4	1	1	0	1

表 2 决策表 DT_2

U_2	a	b	c	d
y_1	0	1	0	0
y_2	1	1	0	1
y_3	1	1	0	1
y_4	0	1	0	0
y_5	1	2	0	0
y_6	1	2	0	1

容易得到 DT_1、DT_2 的属性约简以及 F-并行约简都为 $\{a,b\}$. 假设 $Y_1=\{x:d(x)=0\wedge x\in U_1\}$ 以及 $Y_2=\{x:d(x)=0\wedge x\in U_2\}$,对不在同一个决策子表中的概念集 $S=\{Y_1,Y_2\}$ 进行约简得到 S 的约简也为 $\{a,b\}$.

5 异构数据的属性约简

异构数据(属性集不同)属性约简的目的除了删除冗余属性外,更多地是比较异构数据的异同,抽取出影响概念异同的关键因素. 比如:比较“好人”这个概念在不同时间、空间或条件下的意义. 虽然各种情况下“好人”的概念都有所不同,但都可以转化为两两相互比较,正如多元关系可以转化为若干个二元关系一样,在研究异构数据属性约简和概念漂移的时候往往把注意力放在 2 类不同的数据或概念上.

在两类异构数据或概念的情况下,设异构决策系统簇 $F=F_1\cup F_2$,异构概念集 $S=Y\cup Z$(注:从这里开始 Y 表示概念集). 其中

$F_1=\{DS_i:DS_i=(U_i,A_1,d_1),i=1,2,\cdots,n\}$,

$F_2=\{DS'_j:DS'_j=(U'_j,A_2,d_2),j=1,2,\cdots,m\}$,

$Y=\{Y_1,Y_2,\cdots Y_N\}(i=1,2,\cdots N)$,

$Z=\{Z_1,Z_2,\cdots Z_M\}(j=1,2,\cdots M)$,

$Y\subseteq\cup_{i=1}^{n}U_i/\{d_1\}$,$Z\subseteq\cup_{j=1}^{m}U'_j/\{d_2\}$. 令 $A=A_1\cup A_2$

定理 6 异构决策系统簇 $F=F_1\cup F_2$ 和异构数据概念集 $S=Y\cup Z$. 如果 B 是概念集 S 的约简,则 $B\cap A_1$ 是概念集 Y 约简的超集,$B\cap A_2$ 是概念集 Z 约简的超集.

一般情况下,$B\cap A_1$ 是概念集 Y 的约简,$B\cap A_2$ 是概念集 Z 的约简,但有时候可能会含有冗余属性.

推论 2 设异构决策系统簇 $F=F_1\cup F_2$ 和异构数据概念集 $S=Y\cup Z$. 如果 $A_1\cap A_2=\varnothing$,则属性集 S 的约简等于属性集 Y 的约简与属性集 Z 的约简之并.

接下来,讨论异构数据的属性约简算法. 与结构相同的数据有些不同,对于异构数据的属性约简,可以通过约简的定义,逐一约去不同时在两个数据集中的属性,再约简两个数据集共同的属性. 这个方法与经典的 Pawlak 属性约简算法类似. 由于这个算法不一定能找到最好或者次好的属性约简,一般情况下,可以改用属性重要性的方法求取异构数据的属性约简.

算法 1 异构数据约简算法

输入:异构决策系统簇 $F=F_1\cup F_2$,异构概念集 $S=Y\cup Z$. 设 $A=A_1\cup A_2$,$C=A_1\cap A_2$.

输出:异构概念集 S 的约简 B.

步骤:

1. $B=Core(Y)\cup Core(Z)$;//分别计算概念集 Y 和 Z 的属性核 $Core(Y)$ 和 $Core(Z)$,这两个属性核之并即为异构数据概念集 S 的属性核.

2. 循环执行下列步骤,直到 $POS_A(S)=POS_B(S)$ 或 $C=\varnothing$ 为止:

2.1 对任意的 $\alpha\in C$,计算 $\sigma'(S,B,\alpha)$,如果 $\sigma'(S,B,\alpha)=0$,那么 $C=C-\{\alpha\}$;//计算属性交集 C 中每个元素相对于 B 的外属性重要性

2.2　选取属性重要性最大的属性 α,$B=B\cup\{\alpha\}$,$C=C-\{\alpha\}$//若有多个属性的属性重要性最大,则可以随机选取一个

3. 若 $POS_A(Y)\neq POS_B(Y)$,则执行下列步骤:

3.1　$E=A_1-B-C$;//从概念集 Y 的特有属性中增加属性

3.2　循环执行下列步骤,直到 $POS_A(Y)=POS_B(Y)$ 为止:

3.2.1　对任意的 $\alpha\in E$,计算 $\sigma'(Y,B,\alpha)$,如果 $\sigma'(Y,B,\alpha)=0$,那么 $E=E-\{\alpha\}$;

3.2.2　选取属性重要性最大的属性 α,$B=B\cup\{\alpha\}$,$E=E-\{\alpha\}$;//若有多个属性的属性重要性最大,则可以随机选取一个

4. 若 $POS_A(Z)\neq POS_B(Z)$,则执行下列步骤:

4.1　$E=A_2-B-C$;//从概念集 Z 的特有属性中增加属性

4.2　循环执行下列步骤,直到 $POS_A(Z)=POS_B(Z)$ 为止:

4.2.1　对任意的 $\alpha\in E$,计算 $\sigma'(Z,B,\alpha)$. 如果 $\sigma'(Z,B,\alpha)=0$,那么 $E=E-\{\alpha\}$;

4.2.2　选取属性重要性最大的属性 α,$B=B\cup\{\alpha\}$,$E=E-\{\alpha\}$;//若有多个属性的属性重要性最大,则可以随机选取一个

5. 输出异构概念集 S 的约简 B.

例 2　$DT_3=(U_1,A_1,d)$,$DT_4=(U_2,A_2,d)$是不同地区的流感决策表,如表 3 、表 4 所示. $A_1=\{a,b,c\}$,$A_2=\{a,b,c,e\}$,a 表示头痛,b 表示肌肉痛,c 表示体温,e 表示关节痛,d 表示流感. $V_a=\{0,1\}$,0 表示“否”,1 表示“是”. $V_b=\{0,1\}$,0 表示“否”,1 表示“是”. $V_c=\{0,1,2\}$,0 表示“正常”,1 表示“高”,2 表示“很高”. $V_e=\{0,1\}$,0 表示“否”,1 表示“是”. $V_d=\{0,1\}$,0 表示“否”,1 表示“是”.

这两个异构决策表中关于决策属性 d 有 4 个概念,分别是

$U_1/\{d\}=\{Y_1,Y_2\}=\{\{x_1,x_2,x_3,x_6\},\{x_4,x_5,\}\}$,

$U_2/\{d\}=\{Z_1,Z_2\}=\{\{y_1,y_2\},\{y_3,y_4,y_5,y_6\}\}$,

$A_1\cup A_2=A_2$,$A_1\cap A_2=A_1$. 根据概念约简的定义,可以求取任意单个概念或概念组合的约简. 为简单起见,只求取 $S=\{Y_1,Z_2\}$ 的约简.

首先求 Y_1 和 Z_2 的核属性:$\sigma(Y_1,A_1,a)=0$;

表 3　决策表 DT_3

U_3	a	b	c	d
x_1	0	1	1	1
x_2	1	0	1	1
x_3	1	1	2	1
x_4	0	1	0	0
x_5	1	0	1	0
x_6	0	1	2	1

表 4　决策表 DT_4

U_4	a	b	c	e	d
y_1	0	1	0	0	0
y_2	1	0	0	0	0
y_3	1	0	1	0	1
y_4	1	0	2	1	1
y_5	0	0	2	1	1
y_6	1	0	1	1	1

$\sigma(Y_1,A_1,b)=0$;

$$\sigma(Y_1,A_1,c)=\frac{|POS_{A_1}(Y_1)-POS_{A_2-\{c\}}(Y_1)|}{|Y_1|}=\frac{1}{2}>0$$

同理,

$\sigma(Z_2,A_2,a)=0$,$\sigma(Z_2,A_2,b)=0$,

$\sigma(Z_2,A_2,c)=\frac{1}{4}>0$,$\sigma(Z_2,A_2,e)=0$

所以,Y_1 和 Z_2 的核属性都为$\{c\}$.

接下来求取 S 约简. 令 $B=\{c\}$,$E=(A_1\cap A_2)-B=\{a,b\}$,$\sigma'(S,B,a)=\frac{1}{8}$,$\sigma'(S,B,b)=\frac{1}{8}$. 于是,$B=\{a,c\}$或$B=\{b,c\}$. 易知$B=\{a,c\}$,$B=\{b,c\}$或是异构概念集 S 的约简. Y_1 的属性约简为$\{a,c\}$或$\{b,c\}$,Z_2 的属性约简为$\{c\}$.

6　异构数据的概念漂移探测

不失一般性,可以假设异构数据的决策属性相同. 与结构相同的数据有所不同,异构数据因为其条件属性集可能不同,所以把异构数据之间的概念漂移分为质概念漂移和量概念漂移.

定义 4　决策系统簇 $F=\{DS_1,DS_2\}$ 和异构数据概念集 $S=Y\cup Z$. 其中 $DS_1=(U_1,A_1,d)$,$DS_2=(U_2,A_2,d)$,$Y=\{Y_1,Y_2,\cdots,Y_N\}\subseteq U_1/\{d\}$ $(i=1,2,\cdots,N)$,$Z=\{Z_1,Z_2,\cdots,Z_M\}\subseteq U_2/\{d\}$ $(j=1,2,\cdots,M)$. 设 B 是 S 的属性约简,B_1、B_2 分别是 Y、Z 属性约简,且 $B=B_1\cup B_2$,则 B_1-B_2 称为 Y 相对于 Z 的质概念漂移,B_2-B_1 称为 Z 相对于 Y 的质概念漂移,$(B_1-B_2)\cup(B_2-B_1)$ 称为 S 内的质概念漂移.

定义 5　决策系统簇 $F=\{DS_1,DS_2\}$ 和异构数据概念集 $S=Y\cup Z$. 设 B 是 S 的属性约简,B_1、B_2 分别是 Y、Z 属性约简,且 $B=B_1\cup B_2$,则 Y 和 Z 之间的量概念漂移定义为:$\Delta(S,B)=\sqrt{\sum_{\alpha\in B}(\sigma(Y,B,\alpha)-\sigma(Z,B,\alpha))^2}$,其中

$$\sigma(Y,B,\alpha)=\begin{cases}\sigma(Y,B_1,\alpha), & \text{if } \alpha\in B_1;\\ 0, & \text{else.}\end{cases}$$

$$\sigma(Z,B,\alpha)=\begin{cases}\sigma(Z,B_2,\alpha), & \text{if } \alpha\in B_2;\\ 0, & \text{else.}\end{cases}$$

量概念漂移体现了同一概念在不同时间、空间或条件下属性重要性表示下的欧式距离,是一种数量上的差异,而非本质不同.

下面我们将讨论质概念漂移和量概念漂移的性质:

命题 1　量概念漂移 $\Delta(S,B)$ 非负、对称且满足三角不等式.

命题 2　决策系统簇 $F=\{DS_1,DS_2\}$ 和异构数据概念集 $S=Y\cup Z$. $B\subseteq A$ 是 S 的属性约简,B_1,B_2 分别是 Y

和 Z 属性约简,且 $B=B_1\cup B_2$,对于任意的 $\alpha\in B$ 都有 $\sigma(Y,B,\alpha)=\sigma(Y,B_1,\alpha)$,$\sigma(Z,B,\alpha)=\sigma(Z,B_2,\alpha)$.

命题 3 如果两个异构数据发生了质概念漂移,则量概念漂移一定大于 0.

定理 7 决策系统簇 $F=\{DS_1,DS_2\}$ 和异构数据概念集 $S=Y\cup Z$. $B\subseteq A$ 是 S 的属性约简,B_1、B_2 分别是 Y 和 Z 的属性约简,且 $B=B_1\cup B_2$. 对于任意的 $\alpha\in B$ 都有 $\sigma(S,B,\alpha)\neq 0$ iff $\sigma(Y,B,\alpha)\neq 0$ 或 $\sigma(Z,B,\alpha)\neq 0$.

接下来讨论异构数据概念漂移探测算法.

算法 2 异构数据的概念漂移探测算法

输入:决策系统簇 $F=\{DS_1,DS_2\}$、异构数据概念集 $S=Y\cup Z$.

输出:异构数据集 Y 与 Z 概念漂移判定.

步骤:

1. 调用算法 1 计算 Y 与 Z 的属性约简 $B=B_1\cup B_2$;//B_1 为 Y 的属性约简,B_2 为 Z 的属性约简

2. 计算 Y 与 Z 的质概念漂移 B_1-B_2,B_2-B_1 及 $(B_1-B_2)\cup(B_2-B_1)$;

3. 对于任意 $\alpha\in B_1$ 计算属性重要性 $\sigma(Y,B_1,\alpha)$;

4. 对于任意的 $\alpha\in B_2$ 计算属性重要性 $\sigma(Z,B_2,\alpha)$;

5. 计算 Y 与 Z 的量概念漂移 $\Delta(S,B)$;

6. 输出 Y 与 Z 的质概念漂移 B_1-B_2,B_2-B_1 及 $(B_1-B_2)\cup(B_2-B_1)$;

7. 输出 Y 与 Z 的量概念漂移 $\Delta(S,B)$.

例 3 续例 2. 易知,取概念 Y_1 的约简为 $B_1=\{a,c\}$,概念 Z_2 的约简为 $B_2=\{c\}$,则 S 的质概念漂移为 $(B_1-B_2)\cup(B_2-B_1)=\{a\}$. $\sigma(Y_1,B_1,a)=\frac{1}{4}$;$\sigma(Y_1,B_1,c)=\frac{3}{4}$;$\sigma(Z_2,B_2,c)=1$;于是,量概念漂移 $\Delta(S,B)=\sqrt{\sum_{\alpha\in B}(\sigma(Y_1,B,\alpha)-\sigma(Z_2,B,\alpha))^2}=\frac{\sqrt{2}}{4}$. 从质概念漂移可知,两个表对流感的判定并不相同,表 3 需要属性 $\{a,c\}$ 决定是否流感,而表 4 只需要属性 $\{c\}$ 即可判定流感,它们之间的距离可用量概念漂移来表示. 因此,异构数据之间的概念漂移可以用质概念漂移和量概念漂移来表示.

异构数据概念漂移探测具有很强的理论和现实意义. 在大数据时代,无论电子商务、个性化设计、还是个性化医疗等,在不同时间、空间和条件下都有所不同,通过概念漂移探测研究这些异同可以指导人们更好地工作、学习和生活.

7 结论与展望

本文探究了属性约简的本质,定义了概念(或概念集)的属性约简和属性依赖度、属性重要性,它们兼容值约简、Pawlak 约简和并行约简. 研究了概念属性约简的性质,提出了异构数据的属性约简算法和概念漂移探测算法. 把异构数据的概念漂移分为质概念漂移和量概念漂移,扩展了概念漂移探测的方法. 理论分析和示例显示其有效性. 本文的工作为粗糙集、粒计算进一步融入大数据时代洪流中提供了一种新思路,为大数据和数据流挖掘(特别是异构数据流挖掘)提供了新方法.

进一步研究为:探索异构数据更高效的属性约简和概念漂移探测方法,并把它们运用于电子商务和个性化医疗等实际应用中.

参考文献

[1] Pawlak Z. Rough Sets—Theoretical Aspect of Reasoning about Data[M]. Dordrecht:Kluwer Academic Publishers,1991.

[2] 邓大勇,陈林. 并行约简与 F-粗糙集. 云模型与粒计算[M]. 北京:科学出版社,2012. 210 – 228.
Deng D Y, Chen L. Parallel Reducts and F-rough Sets. Cloud Model and Granular Computing[M]. Beijing: Science Press,2012. 210 – 228. (in Chinese)

[3] Qian J,Miao D Q,Zhang Z H,et al. Parallel reduction algorithm using MapReduce[J]. Information Sciences, 2014, 279:671 – 690.

[4] Wang F,Xu J,Li L. A novel rough set reduct algorithm to feature selection based on artificial fish swarm algorithm[A]. LNCS8795:Proc of 5th International Conference on Swarm Intelligence[C]. Berlin:Springer,2014. 24 – 33.

[5] Liu Y,Huang W L,Jiang Y L,et al. Quick attribute reduct algorithm for neighborhood rough set model[J]. Information Sciences,2014,271:65 – 81.

[6] Hu F,Wang G Y. Knowledge reduction based on divide and conquer method in rough set theory[J]. Mathematical Problems in Engineering,2012,(1):542 – 551.

[7] Eskandari S,Javidi M M. Online streaming feature selection using rough sets[J]. International Journal of Approximate Reasoning,2016,69(C):35 – 57.

[8] Lin T Y,Liu Y,Huang W L. Unifying rough set theories via large scaled granular computing[J]. Fundamenta Informaticae,2013,127:413 – 428.

[9] Cao F Y,Huang J Z. A concept-drfting detection algorithm for categorical evolving data[A]. LNAI 7819:Proc of the 17th Pacific-Asia Conf on Knowledge Discovery and Data Mining[C]. Berlin:Springer,2013. 485 – 496.

[10] 邓大勇,徐小玉,黄厚宽. 基于并行约简的概念漂移探测[J]. 计算机研究与发展,2015,52(5):1071 – 1079.
Deng D Y,Xu X Y,Huang H K. Concept drifting detection for categorical evolving data based on parallel reducts[J]. Journal of Computer Research and Development, 2015,52(5):1071 – 1079. (in Chinese)

[11] 邓大勇,苗夺谦,黄厚宽. 信息表中概念漂移与不确定性分析[J]. 计算研究与发展,2016,53(11):2607 - 2612.
Deng D Y, Miao D Q, Huang H K. Analysis of concept drifting and uncertainty in an information system [J]. Journal of Computer Research and Development, 2016, 53(11): 2607 - 2612. (in Chinese)

[12] 梁吉业,钱宇华,李德玉,等. 大数据挖掘的粒计算理论与方法[J]. 中国科学 E 辑 信息科学,2015,45(11):1355 - 1369.
Liang J Y, Qian Y H, Li D Y, et al. Theory and method of granular computing for big data mining [J]. Science in China Ser E Information Sciences, 2015, 45(11): 1355 - 1369. (in Chinese)

[13] Hu Q, Yu D, Liu J, et al. Neighborhood rough set based heterogeneous feature subset selection [J]. Information Sciences, 2008, 178(18): 3577 - 3594.

[14] Chen D, Yang Y. Attribute reduction for heterogeneous data based on the combination of classical and fuzzy rough set models [J]. IEEE Transactions on Fuzzy Systems, 2014, 22(5): 1325 - 1334.

[15] Qian Y, Liang J, Yao Y, et al. MGRS: A multi-granulation rough set [J]. Information Sciences, 2010, 180(6): 949 - 970.

[16] Lu N, Zhang G, Lu J. Concept drift detection via competence models[J]. Artificial Intelligence, 2014, 209(1): 11 - 28.

[17] Lu N, Lu J, Zhang G. A concept drift-tolerant case-base editing technique [J]. Artificial Intelligence, 2015, 230(C): 108 - 133.

[18] 孙雪,李昆仑,韩蕾,等. 基于特征项分布的信息熵及特征动态加权概念漂移检测模型[J]. 电子学报,2015,43(7):1356 - 1361.
Sun X, Li K L, Han L, et al. Construction of the concept drift detection model based on the information entropy of feature distribution and dynamic weighting algorithm[J]. Acta Electronica Sinica, 2015, 43(7): 1356 - 1361. (in Chinese)

[19] Li P, Wu X, Hu X, Learning concept-drifting data streams with random ensemble decision trees [J]. Neurocomputing, 2015, 166(C): 68 - 83.

作者简介

邓大勇　男,1968 年出生,副教授,博士,现为浙江师范大学行知学院教师. 主要研究方向为粗糙集、粒计算、数据挖掘等.
Email: dayongd@163.com