二进制粒计算模型

郑鹭斌[1] 陈玉明[1,2] 曾志强[1] 卢俊文[1,2]
(厦门理工学院计算机与信息工程学院 厦门 361024)[1]
(江西省高性能计算重点实验室江西师范大学国家网络化支撑软件国际科技合作基地 南昌 330027)[2]

摘 要 粒计算是一种处理不确定性数据的理论方法,涵盖粗糙集、模糊集、商空间、词计算等。目前,数据的粒化与粒的计算主要涉及集合的运算与度量,集合运算的低效制约着粒计算相关算法的应用领域。为此,提出了一种二进制粒计算模型,给出了粒的三层结构,包括粒子、粒群与粒库,并定义了二进制粒子及二进制粒子的运算,将传统的集合运算转化为二进制数的计算,进一步给出了二进制粒子的距离度量,将等价类的集合表示方式转化为粒子的距离度量表示方式,给出了粒子距离的相关性质。该模型定义了二进制粒群距离的概念,给出了二进制粒群距离的计算方法,提出了基于二进制粒群距离的属性约简方法,证明了该方法与经典粗糙集约简方法的等价性,并以二进制粒群距离作为启发式信息,给出了两种约简算法。

关键词 粒计算,粗糙集,二进制粒,粒群距离
中图法分类号 TP181 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.1.058

Binary Granular Computing Model

ZHENG Lu-bin[1] CHEN Yu-ming[1,2] ZENG Zhi-qiang[1] LU Jun-wen[1,2]
(College of Computer and Information Engineering, Xiamen University of Technology, Xiamen 361024, China)[1]
(Key Laboratory of High Performance Computing Center of Jiangxi Province, Nanchang 330027, China)[2]

Abstract Granular computing is a theory dealing with uncertain data, including rough set, fuzzy set, quotient space, computing with words, etc. At present, the granulation of data and granular computing are mainly related to the set operations. As we know, these set operations are inefficient, resulting in restrict the applications of granular computing algorithms. Therefore, we proposed a binary granular computing model, which has the three layer structure including granule, granule swarm and granule library. We defined binary granules and granule operations, which can transform the set operations into the binary number calculations. Furthermore, we proposed a distance metric of two binary granules, which represents the distance of the set of equivalence classes, and discussed some properties of the granule distance. The binary granular computing model defines the concept of binary granule swarm distance, gives the calculation method of binary granule swarm distance, and puts forward the method of attribute reduction based on binary granule swarm distance. We proved the equivalence of our proposed reduction method and the classical Pawlak reduction method. We presented two kinds of reduction algorithm, which use the binary granule swarm distance as the heuristic information.

Keywords Granular computing, Rough sets, Binary granules, Granule swarm distance

1 引言

日益快速的互联网发展背后,积累了大量的、复杂的与不确定性的大数据。人类在认识这些复杂多样与不确定数据的过程中,自然将复杂变为简单,模糊变清晰,大块划分为小粒,形成了一种粒度的思维与计算方法。1965年Zadeh提出了模糊集理论[1],体现了模糊变清晰的粒化思想。1982年Palak提出了粗糙集理论[2],采用等价类划分,体现了知识的粒度化。1985年Hobss提出了粒度[3]的概念,展现了粒的可度量特性。1996年Zadeh提出了信息粒[4]与词计算[5]的思想,1998年T. Y. Lin提出了粒计算[6]的概念,此后,粒计算的研究引起了众多学者的关注与兴趣[7-11]。粒计算是一种处理不确定性数据的理论与方法,横跨多门学科,具有多个分支模型,涵盖粗糙集、模糊集、商空间、词计算等模型[12-16]。

在粒计算的研究中,粒的定义与计算成为粒计算的基础与关键。信息粒是指人们在认识、推理和决策中将大量复杂的信息按照各自的特征和性能划分成若干较简单的块、类、群或组等基本单位[17]。这种基本单位称为粒,这种信息处理的方式称为信息粒化,不同层次信息粒的转化与合并称为粒的计算。不同的抽象层次粒化形成不同的信息粒。粒存在于特

到稿日期:2015-10-11 返修日期:2015-11-11 本文受国家自然科学基金(61573297),福建省自然科学基金(2015J01277),江西师范大学国家网络化支撑软件国际科技合作基地开放课题(NSS1404,NSS1405)资助。

郑鹭斌(1975—),男,硕士,工程师,主要研究方向为粒计算、粗糙集;陈玉明(1977—),男,博士,副教授,主要研究方向为粗糙集、粒计算、数据挖掘;曾志强(1971—),男,博士,副教授,主要研究方向为人工智能、数据挖掘;卢俊文(1981—),男,硕士,实验师,主要研究方向为云计算、数据挖掘。

定的层中，同一层次的粒之间既可以是不相交的，也可以是重叠的，它们相互补充、相互呼应。粒是对一个问题从不同的侧面进行详细的描述。然而，一方面，粒的概念众说纷纭，一个对象是一个粒，一个等价类是一个粒，一个划分是一个粒。对象、等价类与划分是不同层次的描述，不同层次、各种各样粒的定义与描述，使得粒的概念不清晰，造成粒的定义复杂化，反而不利于问题的求解。另一方面，粒计算包括多个分支模型，这些模型下的粒都是一种集合的表示方式，粒的计算也是集合的交、并等运算，集合运算的低效在一定程度上限制了粒计算研究的应用范围。

物以类聚，人以群分。人们在认识、推理与决策的过程中，是基于概念的认知与推理的。一个对象或者物体，我们并不能认识它，只有聚集成一类，抽象成一个概念，才能认识它。因此，概念是认知的最小单位，推理是基于概念的推理。对问题论域的观察可以从不同的侧面进行，从一个观察侧面将论域划分为多个不同的概念，则这多个概念同属于一个观察侧面（划分）。不同的观察侧面则形成多个划分。所有的观察侧面形成的划分组成一个知识的海洋。从人类自然的认识出发，我们将一个等价类定义为一个粒子，一个划分定义为一个粒群，所有的划分组成一个粒库。一个粒子是一个等价类，相当于一个概念，是认知与推理的最小单位。粒计算研究主要关注问题的认知与推理，因而粒子也是粒计算的最小单位。众所周知，计算机系统应用层是一些符号的计算与处理，底层则是二进制数的加、减运算。本文探索一种二进制数的表示方式，等价类表示为一个二进制数，一个粒子则是一个二进制数，从而将粒计算理论中的集合运算转化为二进制数运算。

面对量大、复杂、不确定性的大数据，问题的求解需要采用新的模型与方法。我们从人类认知与推理的自然行为出发，提出一种新的粒计算模型。该模型定义了粒子、粒群、粒库构成的粒结构，给出了粒子间的距离度量与粒群间的距离度量；采用了二进制数的表示方式，将集合运算转化为二进制数运算；定义基于粒群距离的启发式信息，用于粒计算的约简，在粒库中搜索获得最小约简，证明了基于粒群距离的属性约简与传统粗糙集约简的等价性，提出了两种启发式约简方法。

2 粒结构及二进制粒子运算

波兰数学家 Pawlak 提出的粗糙集理论是粒计算研究中采用最为广泛的模型之一。在粗糙集理论中，等价类可以视为一个粒子。任意等价类的交与并得到的等价类也可看成是一个粒子。

定义 1[18] 称 $IS=(U,A,V,f)$ 为信息系统，其中 U 是非空有限集，称为论域，A 是有限属性集，$V=\bigcup_{a\in A}V_a$，V_a 表示属性 a 的值域，$f:U\times A\to V$ 是一个信息函数，即 $\forall x\in U,a\in A$，有 $f(x,a)\in V_a$。任一属性子集 $B\subseteq A$ 决定了一个等价关系 $IND(B)$：$IND(B)=\{(x,y)\in U\times U\mid \forall a\in B,f(x,a)=f(y,a)\}$。

$U/IND(B)$ 构成了 U 的一个划分，划分是等价类的集合。

定义 2 设 U 为一个非空论域，R 为 U 上的一个等价关系，U/R 为 U 关于 R 的划分，$[x]_R\in U/R$ 为 $x\in U$ 的 R 等价类，U 关于 R 的粒子、粒群与粒库定义如下：

1）若 $g(x)=[x]_R$，称 $g(x)$ 为 x 的 R 粒子；

2）若 $G_R=\{g(x)\mid \forall x\in U\}$，称 G_R 为 R 粒群；

3）若 $K_R=\{G_P\mid \forall P\subseteq R\}$，称 K_R 为 R 粒库。

从定义 2 可知，等价类表示成粒子，粒子的集合形成粒群，所有粒群的集合则构成粒库。下面用二进制数来表示粒子。

定义 3 设 U 为一个具有 n 个对象的非空论域，R 为 U 上的一个等价关系，$g(x)$ 为 x 的 R 粒子，称 $b(x)=a^n$ 为 x 的二进制 R 粒子，其中 $a\in\{0,1\}$，$\forall x_i\in U$，如果 $x_i\in g(x)$，则 $a_i=1$，否则 $a_i=0$。

定义 4 设 $b(x)$ 为一个二进制 R 粒子，则该粒子的大小为含有 1 的总数，记为：

$$Size(b(x))=|b(x)|=\sum_{i=1}^{n}a_i$$

0^n 表示空粒子，记为 null，粒子的大小为 0；1^n 表示满粒子，记为 full，大小为。为方便，粒子 $g(x)$ 简记为 g。

定义 5 设 $G_R=\{b(x_i)\mid \forall x_i\in U\}=\{b_1,b_2,\cdots,b_m\}$ 为一个二进制 R 粒群，则该粒群的粒度定义为：

$$GM(G_R)=\frac{1}{|U|^2}\sum_{i=1}^{|U|}|b(x_i)|=\frac{1}{|U|^2}\sum_{i=1}^{m}|b_i|^2$$

定义 6 设 $b=p^n$ 与 $f=q^n$ 为两个二进制粒子，粒子的运算定义如下：

(1)粒子交运算为 $b\wedge f=\{p_i\wedge q_i\}^n$；

(2)粒子并运算为 $b\vee f=\{p_i\vee q_i\}^n$；

(3)粒子减运算为 $b-f=\{p_i-q_i\}^n$；

(4)粒子异或运算为 $b\oplus f=\{p_i\oplus q_i\}^n$；

(5)粒子否运算为 $\neg b=\{\neg p_i\}^n$。

其中，p_i,q_i 分别表示粒子 a,b 中的第 i 位二进制数；$b\wedge f=\{p_i\wedge q_i\}^n$ 表示两个二进制粒子按位相与。

性质 1 根据粒子运算的定义可以得出以下性质：

(1)$b\wedge$ null=null；

(2)$b\wedge$ full=b；

(3)null$\wedge$ full=null；

(4)$b\vee$ null=b；

(5)$b\vee$ full=full；

(6)null$\vee$ full=full；

(7)$b-$null=b；

(8)$b-$full=null；

(9)null$-$full=null；

(10)null$-b$=null；

(11)full$-b=\neg b$；

(12)full$-$null=full；

(13)$\neg$null=full；

(14)$\neg$full=null；

(15)null$\oplus$ null=null；

(16)null$\oplus$ full=full。

例 1 设论域 $U=\{x_1,x_2,x_3,x_4,x_5,x_6,x_7,x_8\}$，划分 $U/R=\{\{x_1,x_2\},\{x_3,x_7,x_8\},\{x_4,x_5,x_6\}\}$，则粒子 $b_1=b_2=11000000=1^20^6$，$b_3=b_7=b_8=00100011=0^21^31^2$，$b_4=b_5=b_6=00011100=0^31^30^2$；粒子的大小 $Size(b_1)=Size(b_2)=2$，$Size(b_3)=Size(b_7)=Size(b_8)=3$，$Size(b_4)=Size(b_5)=$

$Size(b_6)=3$；粒群 $G_R=\{1^20^6,0^21^30^31^2,0^31^30^2\}$，粒群的粒度 $GM(G_R)=(2*2+3*3+3*3)/64=0.3438$；粒子的运算 $b_1\wedge 0^8=0^8$，$b_1\wedge 1^8=b_1$，$b_1\wedge b_2=b_1$，$b_1\wedge b_3=0^8$，$b_1\vee 0^8=b_1$，$b_1\vee 1^8=1^8$，$b_1\vee b_2=b_1$，$b_1\vee b_3=1^30^31^2$，$b_1-b_2=0^8$，$b_1-b_3=1^20^6$，$\neg b_1=0^21^6$，$b_1\oplus b_2=\text{null}$，$b_1\oplus b_3=1^30^31^2$。

3 二进制粒计算的距离度量与约简模型

基于等价关系的信息熵、互信息、知识粒度等概念度量了知识的粗细程度，并反映了决策系统中的分类能力大小，但主要基于等价类的集合运算。本文提出一种新的知识表示方式：二进制粒化的方法，它将等价类的集合运算转化为二进制数的计算，并提出了粒子之间的两种距离度量：粒距离与汉明距离，进一步定义了粒群之间的距离度量，提出基于粒群距离的约简方法，证明了该约简方法与传统粗糙集约简方法的等价性。

3.1 二进制粒计算的距离度量

定义 7 设 $g=p^n$ 与 $f=q^n$ 为两个二进制粒子，两个粒子的粒距离定义为：

$$d(g,f)=\frac{|g\oplus f|}{|g\vee f|}=\frac{\sum_{i=1}^{n}p_i\oplus q_i}{\sum_{i=1}^{n}p_i\vee q_i}$$，其中 $g\vee f\neq\text{null}$。

定理 1 两个二进制粒子的粒距离是一种距离度量，满足以下 3 个性质：

(1)非负，$d(g,f)\geqslant 0$；

(2)对称，$d(g,f)=d(f,g)$；

(3)三角不等式，$d(g,k)+d(k,f)\geqslant d(g,f)$。

证明：(1)和(2)两条性质根据粒距离的定义很容易得证，证明略。下面证明性质(3)。

给定任意 a,b,c 3 个数，其中 $a\geqslant b>0,c\geqslant 0$，则不等式 $\frac{b}{a}\leqslant\frac{b+c}{a+c}$ 成立。根据粒距离的定义，则

$$\begin{aligned}
&d(g,k)+d(k,f)-d(g,f)\\
&=\frac{|g\vee k|-|g\wedge k|}{|g\vee k|}+\frac{|k\vee f|-|k\wedge f|}{|k\vee f|}-\frac{|g\vee f|-|g\wedge f|}{|g\vee f|}\\
&=1-\frac{|g\wedge k|}{|g\vee k|}-\frac{|k\wedge f|}{|k\vee f|}+\frac{|g\wedge f|}{|g\vee f|}\\
&=1-\frac{|g\wedge k|+(|f|-|f\wedge(g\vee k)|)}{|g\vee k|+(|f|-|f\wedge(g\vee k)|)}-\frac{|k\wedge f|+(|g|-|g\wedge(k\vee f)|)}{|k\vee f|+(|g|-|g\wedge(k\vee f)|)}+\frac{|g\wedge f|}{|g\vee k\vee f|}\\
&=1-\frac{|g\wedge k|+|f|-(|g\wedge f|+|k\wedge f|-|g\wedge k\wedge f|)}{|g\vee k\vee f|}-\frac{|k\wedge f|+|g|-(|g\wedge k|+|g\wedge f|)}{|g\vee k\vee f|}+\frac{|g\wedge f|}{|g\vee k\vee f|}\\
&=1-\frac{|g|+|f|-|g\wedge f|}{|g\vee k\vee f|}+\frac{2|g\wedge f|-|g\wedge k\wedge f|}{|g\vee k\vee f|}\geqslant 0
\end{aligned}$$

因此，$d(g,k)+d(k,f)-d(g,f)\geqslant 0$，即 $d(g,k)+d(k,f)\geqslant d(g,f)$ 成立。命题得证。

定义 8 设 $g=p^n$ 与 $f=q^n$ 为两个二进制粒子，两个粒子的汉明距离定义为：

$$h(g,f)=|g\oplus f|=\sum_{i=1}^{n}p_i\oplus q_i$$

定理 2 两个二进制粒子的粒距离与汉明距离满足：

$$d(g,f)=\frac{h(g,f)}{|g\vee f|}=\frac{h(g,f)}{h(g\vee f,\text{null})}$$

证明：根据粒距离与汉明距离的定义很容易得证，证明略。

定义 9 设 U 为论域，F 为二进制粒库，S 和 T 为论域 U 上的等价关系，二进制粒群 $G_S,G_T\in F$，$G_S=\{s(x_i)\mid x_i\in U\}$，$G_T=\{t(x_i)\mid x_i\in U\}$，则两个粒群的距离定义为：

$$H(G_S,G_T)=\frac{1}{|U|^2}\sum_{i}^{|U|}h(s(x_i),t(x_i))$$

定理 3 任意两个二进制粒群的距离满足：

$$0\leqslant H(G_S,G_T)\leqslant 1-\frac{1}{|U|}$$

证明：由 $1\leqslant|s(x_i)\wedge t(x_i)|\leqslant|U|$，$1\leqslant|s(x_i)\vee t(x_i)|\leqslant|U|$，可知 $0\leqslant|s(x_i)\vee t(x_i)-s(x_i)\wedge t(x_i)|\leqslant|U|-1$。由此，得到 $0\leqslant|s(x_i)\oplus t(x_i)|\leqslant|U|-1$，从而 $0\leqslant\frac{1}{|U|^2}\sum_{i}^{|U|}|s(x_i)\oplus t(x_i)|\leqslant 1-\frac{1}{|U|}$。根据粒群距离的定义，可知 $0\leqslant H(G_S,G_T)\leqslant 1-\frac{1}{|U|}$。命题得证。

定理 4 两个二进制粒群的距离是一种距离度量，满足以下 3 个性质：

(1)非负，$H(G_S,G_T)\geqslant 0$；

(2)对称，$H(G_S,G_T)=H(G_T,G_S)$；

(3)三角不等式，$H(G_S,G_R)+H(G_R,G_T)\geqslant H(G_S,G_T)$。

证明：(1)和(2)两条性质根据粒群距离的定义很容易得证，证明略。下面证明性质(3)。

根据两个二进制粒汉明距离的定义，可知 $h(s(x_i),r(x_i))+h(r(x_i),t(x_i))\geqslant h(s(x_i),t(x_i))$。由此，$H(G_S,G_R)+H(G_R,G_T)=\frac{1}{|U|^2}\sum_{i}^{|U|}h(s(x_i),r(x_i))+\frac{1}{|U|^2}\sum_{i}^{|U|}h(r(x_i),t(x_i))=\frac{1}{|U|^2}\sum_{i}^{|U|}(h(s(x_i),r(x_i))+h(r(x_i),t(x_i)))\geqslant\frac{1}{|U|^2}\sum_{i}^{|U|}h(s(x_i),t(x_i))=H(G_S,G_T)$。因此，$H(G_S,G_R)+H(G_R,G_T)\geqslant H(G_S,G_T)$ 成立。

定理 5 设 U 为论域，S 为论域 U 上的等价关系，F 为二进制粒库，G_S 为二进制粒群，$\text{null}\in F$，$G_S=\{s(x_i)\mid x_i\in U\}$，$\text{null}=0^n$，则有 $GM(G_S)=H(G_S,\text{null})$。

证明：根据粒群距离和粒子汉明距离的定义，有 $H(G_S,\text{null})=\frac{1}{|U|^2}\sum_{i}^{|U|}h(s(x_i),\text{null})=\frac{1}{|U|^2}\sum_{i}^{|U|}|s(x_i)\oplus 0^n|=\frac{1}{|U|^2}\sum_{i}^{|U|}|s(x_i)|$；令 $G_S=\{s(x_i)\mid x_i\in U\}=\{s_1,s_2,\cdots,s_m\}$，则可知 $\sum_{i}^{|U|}|s(x_i)|=\sum_{i}^{m}|s_i|^2$；因此，$\frac{1}{|U|^2}\sum_{i}^{|U|}|s(x_i)|=\frac{1}{|U|^2}\sum_{i}^{m}|s_i|^2$ 成立；根据粒群的粒度定义，可知 $H(G_S,\text{null})=GM(G_S)$。命题得证。

3.2 二进制粒计算的属性约简模型

在决策系统中，传统粗糙集约简模型是基于正域的约简模型，其特点是保持正域不变，逐步删除一些冗余的属性。在信息系统中，约简模型的特点是保持依赖度或者划分(等价关系)不变，逐步删除冗余的属性。

定义 10[17] 设 $IS=(U,A,V,f)$ 是一个信息系统，$\forall a\in A$，如果 $IND(A-\{a\})=IND(A)$，则称 a 是 A 中不必要的

(冗余的)属性;否则,称 a 是 A 中必要的属性。

定义 11[17] 设 $IS=(U,A,V,f)$ 是一个信息系统,$B\subseteq A$ 是一个属性子集,如果满足

(1)$IND(B)=IND(A)$;

(2)对 $\forall b\in B$,有 $IND(B-\{b\})\neq IND(B)$;

则称 B 是 A 的一个约简。

定理 6 设 $IS=(U,A,V,f)$ 是一个信息系统,$B,C\subseteq A$ 是两个属性子集,G_B,G_C 为二进制粒群,若 $IND(B)=IND(C)$,则粒群距离 $H(G_B,G_C)=0$。

证明:令 $U/B=G_B=\{X_1,X_2,\cdots,X_m\}$,$U/C=G_C=\{Y_1,Y_2,\cdots,Y_n\}$,已知 $IND(B)=IND(C)$,所以 $U/B=U/C$,即 $\{X_1,X_2,\cdots,X_m\}=\{Y_1,Y_2,\cdots,Y_n\}$,则 $\forall X_i\in G_B$,存在 $Y_j\in G_C$ 满足 $X_i=Y_j$,可知 $|X_i\oplus Y_j|=0$。由粒群的距离定义可知,$H(G_B,G_C)=\frac{1}{|U|^2}\sum_{i=1}^{|U|}|X_i\oplus Y_j|$,而 $|X_i\oplus Y_j|=0$,所以 $H(G_B,G_C)=0$。

定理 7 设 $IS=(U,A,V,f)$ 是一个信息系统,$B,C\subseteq A$ 是两个属性子集,且 $B\subseteq C$,G_B,G_C 为二进制粒群,若粒群距离 $H(G_B,G_C)=0$,则 $IND(B)=IND(C)$。

证明:反证法。假设 $IND(B)=IND(C)$ 不成立,即 $IND(B)\neq IND(C)$;由 $B\subseteq C$,知 $IND(B)\supseteq IND(C)$,因为 $IND(B)\neq IND(C)$,所以 $IND(B)\supset IND(C)$;令 $U/B=G_B=\{X_1,X_2,\cdots,X_m\}$,$U/C=G_C=\{Y_1,Y_2,\cdots,Y_n\}$,由 $IND(B)\supset IND(C)$,知 $\{X_1,X_2,\cdots,X_m\}\supset\{Y_1,Y_2,\cdots,Y_n\}$,所以 $\exists X_i\in G_B,Y_j\in G_C$ 满足 $X_i\supset Y_j$,可知 $|X_i\oplus Y_j|\neq 0$。由粒群的距离定义可知,$H(G_B,G_C)=\frac{1}{|U|^2}\sum_{i=1}^{|U|}|X_i\oplus Y_j|$,而 $|X_i\oplus Y_j|\neq 0$,所以 $H(G_B,G_C)\neq 0$,这和已知 $H(G_B,G_C)=0$ 相矛盾。故命题得证。

定理 8 设 $IS=(U,A,V,f)$ 是一个信息系统,G_A 和 $G_{A-\{a\}}$ 为二进制粒群,$\forall a\in A$ 在 A 中是不必要的(冗余的)属性,其充分必要条件是粒群距离 $H(G_{A-\{a\}},G_A)=0$。

证明:(必要性)设 $\forall a\in A$ 在 A 中是不必要的属性,由定义 10 知 $IND(A-\{a\})=IND(A)$ 成立;由定理 6 可知,$H(G_{A-\{a\}},G_A)=0$。

(充分性)设 $\forall a\in A$,由 $H(G_{A-\{a\}},G_A)=0$ 和定理 7 可知,$IND(A-\{a\})=IND(A)$ 成立,故 $\forall a\in A$ 在 A 中是不必要的属性。

推论 1 $\forall a\in A$ 在 A 中是必要的属性,即 $IND(A-\{a\})\neq IND(A)$,其充分必要条件是 $H(G_{A-\{a\}},G_A)\neq 0$。

定理 9 设 $IS=(U,A,V,f)$ 是一个信息系统,$B\subseteq A$,G_A,G_B 为二进制粒群,B 是 A 的一个约简的充分必要条件为:

1)$H(G_B,G_A)=0$;

2)对 $\forall b\in B$,有 $H(G_{B-\{b\}},G_B)\neq 0$。

证明:由定义 11、定理 7、定理 8 及推论 1 可以证明。证明略。

定义 12 设 $IS=(U,A,V,f)$ 是一个信息系统,$G_{A-\{a\}}$,G_A 为粒群,$\forall a\in A$,定义 a 在信息系统中的属性重要度为:

$$Sign(a,A)=H(G_{A-\{a\}},G_A)$$

该属性重要度采用了粒群的距离度量作为衡量标准,单个属性增减后的粒群距离大小的变化反映了该单个属性的重要性程度。因此,该属性重要度可以作为属性约简算法当中的属性选择标准。

3.3 基于二进制粒群距离的属性约简算法

依据定理 9 可知,基于粒群距离的约简模型与传统粗糙集的约简模型是等价的。根据基于粒群距离的属性重要度定义,可设计如下基于粒群距离的属性约简算法。下面以定义 12 的属性重要性为启发式信息设计了一个自底向上的属性约简算法。

算法 1 自底向上方式基于二进制粒群距离的属性约简

输入:信息系统 IS=(U,C,V,f)

输出:IS 的一个约简 R

步骤 1:初始化 R:=∅

步骤 2:对 C−R 重复

(1)如果 $H(G_R,G_C)$ 结束,转步骤 3,否则转(2);

(2)对每个 a∈C−R 计算 $H(G_{R\cup\{a\}},G_R)$;

(3)选择属性 a 满足 $\max_{a\in C-R}\{H(G_{R\cup\{a\}},G_R)\}$;

(4)R:=R∪{a}。

步骤 3:输出 R

下面从整个条件属性集开始,设计了一个自顶向下的约简算法。在算法中,每次选择粒群距离变化最大的属性,若去掉它后形成的粒群与初始粒群的距离为 0,则可以去掉它,否则保留下来,依次进行下去,直到得到一个条件属性子集,在其中去掉任何一个属性,形成的粒群与初始的粒群距离均不为 0,则算法结束,该属性子集即为所求的约简。

算法 2 自顶向下方式基于二进制粒群距离的属性约简

输入:信息系统 IS=(U,C,V,f)

输出:IS 的一个约简 R

步骤 1:R:=C,A:=C

步骤 2:当 A≠∅ 重复

(1)在 A 中选择属性 a 满足 $\max_{a\in R}\{H(G_{R-\{a\}},G_R)\}$;

(2)A:=A−{a};

(3)如果 $H(G_{R-\{a\}},G_C)=0$,则 R:=R−{a}。

步骤 3:输出 R

结束语 通过分析人类认知的特点,提出了粒计算的三层粒结构模型:粒子、粒群与粒库。在信息系统中等价类集合表示的基础上,定义了粒子的二进制表示与运算算子,提出了二进制粒子的两种距离度量方法:粒距离度量与汉明距离度量。针对二进制粒群的特点,定义了粒群距离,提出了基于二进制粒群距离的属性约简模型,证明了该约简模型与传统粗糙集约简模型的等价性。基于此,设计了两种基于二进制粒群距离的属性约简算法。

然而,基于二进制粒群的属性约简方法只适用于等价类划分的信息系统,并没有考虑具有连续性数值的信息系统。针对连续性数值的信息系统,可以采用邻域关系或者覆盖粗糙集的理论进行研究。今后的研究中,将着重针对连续性数值的信息系统约简问题,把二进制粒表示与邻域关系结合起来进行研究。

参 考 文 献

[1] Zadeh L A. Fuzzy sets[J]. Information and Control,1965,8(3):338-353

[2] Pawlak Z. Rough sets[J]. International Journal of Information and Computer Sciences,1982,11(1):341-356

[3] Hobbs J R. Granularity[C]//Proc. of the IJCAI. 1985:432-435
[4] Zadeh L A. Fuzzy logic-computing with words[J]. IEEE Trans. on Fuzzy Systems,1996,4(2):103-111
[5] Zadeh L A. Towards a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic[J]. Fuzzy Sets and Systems,1997,90(2):111-127
[6] Lin T Y. Granular computing on binary relations II:Rough set representations and belief functions[M]//Skowron A,Polkowski L, eds. Rough Sets in Knowledge Discovery. Heidelberg: Physica-Verlag,1998:121-140
[7] Gacek A. Signal processing and time series description:A Perspective of Computational Intelligence and Granular Computing [J]. Applied Soft Computing,2015,27:590-601
[8] Li Jin-hai, Mei Chang-lin, Xu Wei-hua, et al. Concept learning via granular computing: A cognitive viewpoint[J]. Information Sciences,2015,298:447-467
[9] Qian Yu-hua, Zhang Hu, Li Fei-jiang, et al. Set-based granular computing:A lattice model[J]. International Journal of Approximate Reasoning,2014,55(3):834-852
[10] Chen Y M,Miao D Q,Wang R Z. A rough set approach to feature selection based on ant colony optimization[J]. Pattern Recognition Letters,2010,31(3):226-233
[11] An J J,Wang G Y,et al. A rule generation algorithm based on granular computing[C]//2005 IEEE International Conference on Granular Computing. 2005:102-107
[12] Zhang L,Zhang B. Fuzzy reasoning model under quotient space structure[J]. Information Sciences,2005,173(4):353-364
[13] Zhu W, Wang F. Reduction and axiomization of covering generalized rough sets[J]. Information Sciences,2003,152(1):217-230
[14] Zadeh L. A. Fuzzy sets and information granularity[M]//Gupta M,Ragade R,Yager R,eds. Advances in Fuzzy Set Theory and Applications. North-Holland,Amsterdam,1979:3-18
[15] Hu Q H,Yu D R,Xie Z X. Neighborhood classifiers[J]. Expert Systems with Applications,2008,34(2):866-876
[16] Hu Q H,Pedrycz W,Yu D R,et al. Selecting discrete and continuous features based on neighborhood decision error minimization[J]. IEEE Transactions on Systems,Man,and Cybernetics—Part B:Cybernetics,2010,40(1):137-50
[17] Liu Qing. Rough sets and rough reasoning [M]. Beijing:Science Press,2001(in Chinese)
刘清. Rough 集及 Rough 推理[M]. 北京:科学出版社,2001
[18] Miao Duo-qian, Fan Shi-dong. The Calculation of Knowledge Granulation and Its Application [J]. Systems Engineering-theory & Practice,2002,22(1):48-56(in Chinese)
苗夺谦,范世栋. 知识的粒度计算及其应用[J]. 系统工程理论与实践,2002,22(1):48-56
[19] Gou Guang-lei, Huang Li-feng, Ni wei. Conceptual Clustering Algorithm Based on Granular Computing [J]. Journal of Chongqing University of Technology(Natural Science),2013,27(6):76-79(in Chinese)
苟光磊,黄丽丰,倪伟. 基于粒计算的概念聚类算法[J]. 重庆理工大学学报(自然科学),2013,27(6):76-79

(上接第 254 页)

(Random Forest)进行分类。由于时空数据的特殊性(共现的多样性和共现的偶然性),我们提出了在算法中加入位置权重动态调整学习率。实验结果表明 Location-weight 算法提高了用户社交联系强度预测的结果。进一步的研究工作将会围绕时空数据中共现的多样性对用户间社交联系强度预测的影响来展开。

参 考 文 献

[1] Pham H,Shahabi C,Liu Y. Ebm:an entropy-based model to infer social strength from spatiotemporal data[C]//Proceedings of the 2013 International Conference on Management of Data. ACM,2013:265-276
[2] Gonzalez M C,Hidalgo C A,Barabasi A L. Understanding individual human mobility patterns[J]. Nature,2008,453(7196):779-782
[3] Cheng Z, Caverlee J, Lee K, et al. Exploring Millions of Footprints in Location Sharing Services[C]//ICWSM. 2011:81-88
[4] Lindqvist J,Cranshaw J,Wiese J,et al. I'm the mayor of my house:examining why people use foursquare-a social-driven location sharing application[C]//Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM,2011:2409-2418
[5] Kostakos V, Venkatanathan J, Reynolds B, et al. Who's your best friend-targeted privacy attacks In location-sharing social networks[C]//Proceedings of the 13th International Conference on Ubiquitous Computing. ACM,2011:177-186
[6] Crandall D J,Backstrom L,Cosley D,et al. Inferring social ties from geographic coincidences[J]. Proceedings of the National Academy of Sciences,2010,107(52):22436-22441
[7] Cranshaw J, Toch E, Hong J, et al. Bridging the gap between physical location and online social networks[C]//Proceedings of the 12th ACM International Conference on Ubiquitous Computing. ACM,2010:119-128
[8] Li Q,Zheng Y,Xie X,et al. Mining user similarity based on location history[C]//Proceedings of the 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. ACM,2008:34
[9] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[J]. arXiv preprint arXiv:1301.3781,2013
[10] Li Min, Wang Xiao-cong, Zhang Jun, et al. Study on Check-in and Related Behaviors of Location-based Social Network Users [J]. Computer Science,2013,40(10):72-76(in Chinese)
李敏,王晓聪,张军,等. 基于位置的社交网络用户签到及相关行为研究[J]. 计算机科学,2013,40(10):72-76
[11] Mikolov T,Yih W,Zweig G. Linguistic Regularities in Continuous Space Word Representations[C]// HLT-NAACL. 2013:746-751
[12] Mikolov T,Kombrink S,Burget L,et al. Extensions of recurrent neural network language model[C]//2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE,2011:5528-5531