

相关集的数据关联描述及实例讨论

闫林¹ 阮宁¹ 闫硕² 高伟¹

(河南师范大学计算机与信息工程学院 新乡 453007)¹

(北京交通大学计算机与信息技术学院 北京 100044)²

摘要 为了讨论数据关联问题,按层次对数据集进行了粒化,引入了每一粒的相关集,产生了与相关集类关联的分层粒化结构,并称之为粒化树。进而以同一数据集上的两棵粒化树为结构支撑,完成了对数据关联的定义,使相关集之间数据的关联得以数值化表示,形成了数据关联的数值描述方法。对此的研究确定了数据关联的等价条件,以此为依托并通过实例探究了数据关联的相关性质,讨论了关联的紧密程度、数据的粒化等同、关联的相互比较等数值化的处理方法。同时为实例的讨论提供了算法编程的基础,表明了数据关联研究的实际意义。

关键词 划分,粒化树,数据关联,粒,相关集

中图分类号 TP181 文献标识码 A DOI 10.11896/j.issn.1002-137X.2017.01.052

Description of Data Association Occurring in Related Sets and Specific Example

YAN Lin¹ RUAN Ning¹ YAN Shuo² GAO Wei¹

(College of Computer and Information Engineering, Henan Normal University, Xinxiang 453007, China)¹

(School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China)²

Abstract In order to discuss the problem on the data association, a data set was divided into granules according to different levels, and each granule was assigned related set. This gives rise to a hierarchy structure called granulation tree which is associated with classes of related sets. Then, supported by two granulation trees based on the same data set, a definition of the data association was introduced, demonstrating a numerical representation of the data association occurring in related sets, and leading to a way of numerical description on it. The research on the topic brings about a condition that is equivalent to the definition. Also by using the condition and based on a specific example, some properties were investigated, which focuses on numerical analyses of some issues, such as the close degree of the data association, the granulation identity between data, the numerical comparison of data associations, and so on. Accordingly, the discussion on the specific example provides the basis of algorithm programming and shows the practical significance of the research.

Keywords Partition, Granulation tree, Data association, Granule, Related set

1 引言

数据问题的研究常常涉及数据之间的联系,故而出现了数据联系描述问题。这里的数据联系是指数据类之间通过公共数据产生的关联关系,称为数据关联。数据关联的课题源于与数据处理相关的理论和应用研究,更与实际中的关联现象相联系。就数据研究而言,数据挖掘^[1,2]、数据仓储^[3,4]、数据推理^[5-7]、数据分类^[8,9]等信息科学的分支均涉及数据关联的问题。比如挖掘数据与挖掘环境的关系,存储数据之间的依存信息,结论数据对前提数据的依赖,数据分类后的数值化处理等都蕴含着数据之间的联系。另一方面,若观察周围,则可看到许多实际中数据与数据之间通过公共数据产生关联的相关实例:

(1)大学生可使家乡人民与其同学之间产生关联。

(2)航天飞机可使航天器与普通飞机之间产生关联。

(3)张三可使单位同事与家庭成员之间产生关联。

(4)两栖动物可使陆地动物和海洋动物之间产生关联。

(5)间谍可使对立双方之间产生关联。

如果把上述各例中的对象看做数据,那么它们展示的数据关联呈现出这样的模式:公共数据促成的两类数据之间产生关联。此时,一数据类中的数据与另一数据类中的数据通过公共数据关联在了一起。虽然许多研究分支涉及了数据关联问题,但以该模式为主题的关联探究相对较少或不曾有,这为研究工作预留了课题。下面所述的数据关联即指该模式下的关联形式。由于数据关联常存在于实际当中,因此建立对应的描述方法将具有理论和应用方面的意义。

描述方法的产生不仅将拓展理论探究的层面,而且对于问题建模、算法模拟、程序设计和自动管理等也具有实际意义。

到稿日期:2015-11-27 返修日期:2016-03-30 本文受国家自然科学基金项目(U1204606)资助。

闫林(1957—),男,教授,硕士生导师,主要研究方向为数理逻辑、粒计算、数据推理;阮宁(1988—),男,硕士生,主要研究方向为粗糙集、数据挖掘, E-mail: 393789734@qq.com;闫硕(1987—),男,博士生,主要研究方向为计算机代数、数据推理;高伟(1990—),男,硕士生,主要研究方向为粗糙集、数据挖掘。

下文将围绕数据关联问题展开讨论,通过构建数学结构支撑描述方法的产生,并使数学结构融入到方法中。

实际上,文献[10,11]已对数据关联问题进行了探究,为相关的工作提供了研究的基础。其中文献[10]通过关联链的分析确立了数据关联的研究途径,其重要之处在于利用粒计算^[12,13]的数据处理方法构造了称为粒化树的数学结构,为文献[11]的研究做了铺垫。粒化树是文献[11]中数据关联定义的基础,粒化树的层次信息促成了数据信息的数值表示,使数值的变化与数据关联的紧密程度联系在一起。不过文献[10,11]的内容并未涵盖所有问题,文献[10]中涉及的关联链仅关注了粒包含形成的链,聚焦在粒的层面,未考虑数据关联的数学表示;文献[11]虽给出了数据关联的定义,但该定义基于不同数据集上的粒化树,未对同一数据集上相异粒化树的情况进行处理。

因此,针对数据关联的研究,存在着有待探究的方面。基于同一数据集的不同粒化树构建,以及由此支撑的数据关联形式预留了继续探究的空间,下文的讨论将围绕于此。同时本文粒化树的构造将涉及相关集,相关集是数据类的数学表示,是已有研究不曾面对的问题。

2 划分与相关集类

数据关联的讨论离不开数据信息及数据的性质,所以讨论从数据开始。设 U 是一个集合,称为数据集, U 中的元素称为数据。数据是各类对象的抽象,既可以表示人员、物体、企业等实际对象,也可以表示数字、字母、函数等抽象概念。数据集是一类数据的聚集,且常常需要根据数据的性质对数据集实施分类处理。

定义 1^[12] 设 U 是一数据集,如果存在 U 的子集的集合 $A = \{A_1, A_2, \dots, A_k\} (k \geq 1)$, 其中 $A_i \subseteq U (i=1, 2, \dots, k)$, 且满足下述 3 个条件,则称 $A = \{A_1, A_2, \dots, A_k\}$ 是 U 的划分:

- (1) $A_i \neq \emptyset (i=1, 2, \dots, k)$;
- (2) $A_i \cap A_j = \emptyset (i \neq j \text{ 且 } 1 \leq i, j \leq k)$;
- (3) $A_1 \cup A_2 \cup \dots \cup A_k = U$ 。

此外,当 $A_i \in A (i=1, 2, \dots, k)$ 时,称 A_i 为粒。

划分 $A = \{A_1, A_2, \dots, A_k\}$ 中互不相交的粒 $A_1, A_2, \dots, A_k$, 展示了数据集 U 中数据的一种分类,这种分类往往基于数据的某种性质,不妨考虑如下的例子。

例 1 设 $U = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$ 。性质“除以 3 余数相同的数”可以把 U 中的数据予以分类,从而得到划分 $A = \{\{1, 4, 7, 10\}, \{2, 5, 8\}, \{3, 6, 9\}\}$, 其中粒 $\{1, 4, 7, 10\}$ 中每一个数除以 3 的余数是 1, 粒 $\{2, 5, 8\}$ 中每一个数除以 3 的余数是 2, 粒 $\{3, 6, 9\}$ 中每一个数除以 3 的余数是 0。因此该划分由性质“除以 3 余数相同的数”确定产生。

例 2 设 $U = \{x \mid x \text{ 是 79 次列车上的乘客}\}$ 。性质“同一车站上 79 次列车的旅客”可确定 U 的划分 $A = \{A_1, A_2, \dots, A_k\}$, 其中 $A_i \subseteq U (i=1, 2, \dots, k)$, 粒 A_1 表示从车站 1 上 79 次列车旅客的集合, 粒 A_2 表示从车站 2 上 79 次列车旅客的集合, ..., 粒 A_k 表示从车站 k 上 79 次列车旅客的集合。因此性质“同一车站上 79 次列车的旅客”确定了 U 的一种划分。

为了研究数据关联问题,可对划分 $A = \{A_1, A_2, \dots, A_k\}$ 中每一粒 $A_i (i=1, 2, \dots, k)$ 指定一集合 A_i' , 称其为 A_i 的相关集,使得 $A_i \subseteq A_i'$, 同时 A_i' 与 A_i 中的数据存在着相同的性质。

不妨通过例 1 和例 2 的划分进行说明:

(1) 考虑例 1 中的划分 $A = \{\{1, 4, 7, 10\}, \{2, 5, 8\}, \{3, 6, 9\}\}$ 。对于粒 $\{1, 4, 7, 10\}$, 指定 $\{1, 4, 7, 10, 13, 16\}$ 为其相关集,此时 $\{1, 4, 7, 10\} \subseteq \{1, 4, 7, 10, 13, 16\}$, 且 $\{1, 4, 7, 10, 13, 16\}$ 中的数除以 3 的余数也都是 1。对于粒 $\{2, 5, 8\}$, 指定 $\{2, 5, 8, 11, 14, 17\}$ 为其相关集,显然 $\{2, 5, 8\} \subseteq \{2, 5, 8, 11, 14, 17\}$, 且 $\{2, 5, 8, 11, 14, 17\}$ 中的数除以 3 的余数都是 2。对于粒 $\{3, 6, 9\}$, 指定 $\{3, 6, 9, 12, 15\}$ 为其相关集,显然 $\{3, 6, 9\} \subseteq \{3, 6, 9, 12, 15\}$, 且 $\{3, 6, 9, 12, 15\}$ 中的数除以 3 的余数都是 0。于是粒中数据与其相关集中数据具有相同的性质。

(2) 考虑例 2 中的划分 $A = \{A_1, A_2, \dots, A_k\}$, 粒 $A_i (i=1, 2, \dots, k)$ 表示从车站 i 上 79 次列车旅客的集合。 A_i 的相关集 A_i' 可指定为从车站 i 上车(不仅 79 次列车)的所有旅客的集合,此时 $A_i \subseteq A_i'$, 且 A_i 与 A_i' 中的数据都具有“从车站 i 上车”的性质。

实际上,指定 A_i' 为 A_i 的相关集时,数据的多少与指定密切相关。比如上述(1)中,指定 $\{1, 4, 7, 10, 13, 16\}$ 为 $\{1, 4, 7, 10\}$ 的相关集。实际上, $\{1, 4, 7, 10, 13, 16, 19, 22, 25\}$ 也可指定为 $\{1, 4, 7, 10\}$ 的相关集,此时 $\{1, 4, 7, 10, 13, 16, 19, 22, 25\}$ 的每一个数除以 3 的余数都为 1。又如上述(2)中,指定 $A_i' = \{x \mid x \text{ 是从车站 } i \text{ 上车的旅客}\}$ 为 A_i 的相关集,由于 A_i' 涉及了更多的乘客,因此 $A_i \subseteq A_i'$ 。也可以指定 $A_i'' = \{x \mid x \text{ 是从车站 } i \text{ 上 79 次和 80 次列车的旅客}\}$, 此时有 $A_i \subseteq A_i'' \subseteq A_i'$ 。

因此指定粒的相关集将涉及数据的选定范围,由讨论者根据情况确定。特别当涉及具体问题时,粒的相关集将与问题联系在一起,第 5 节的实例将给予展示。

定义 2 设 $A = \{A_1, A_2, \dots, A_k\}$ 是数据集 U 的划分,如果对于每一粒 $A_i \in A (i=1, 2, \dots, k)$, 均存在数据的集合 A_i' , 满足 $A_i \subseteq A_i'$, 且当 $A_i, A_j \in A$ 及 $A_i \neq A_j$ 时,有 $A_i' \cap A_j' = \emptyset$, 则称集合 $A' = \{A_1', A_2', \dots, A_k'\}$ 为划分 A 的相关集类,此时 A' 中的 A_i' 称为粒 $A_i (i=1, 2, \dots, k)$ 的相关集。

当 $A' = \{A_1', A_2', \dots, A_k'\}$ 是划分 $A = \{A_1, A_2, \dots, A_k\}$ 的相关集类时,对于粒 $A_i, A_j \in A$, 若 A_i 与 A_j 不同,则由定义 1 (2)可知 $A_i \cap A_j = \emptyset$, 此时 A_i 和 A_j 的相关集 $A_i' (\in A')$ 和 $A_j' (\in A')$ 也满足 $A_i' \cap A_j' = \emptyset$ 。当 $A_i \cap A_j = \emptyset$ 时,对于 $A_i' \cap A_j' = \emptyset$ 的要求是讨论的需要,上述的(1)和(2)表明了该要求易于满足。

当 $A' = \{A_1', A_2', \dots, A_k'\}$ 是划分 $A = \{A_1, A_2, \dots, A_k\}$ 的相关集类时,对于粒 $A_i \in A$ 和其相关集 $A_i' \in A'$, A_i 和 A_i' 之间满足 $A_i \subseteq A_i'$, 上述的(1)和(2)展示了由粒 A_i 确定其相关集 A_i' 的过程,并且可看出由 A_i 得到 A_i' 往往依据情况而定。

另外,从划分 A 与其相关集类 A' 的定义可知, A 中的粒 A_i 与 A' 中的相关集 A_i' 之间一一对应。

当 $A = \{A_1, A_2, \dots, A_k\}$ 是数据集 U 的划分, $A' = \{A_1', A_2', \dots, A_k'\}$ 是 A 的相关集类时,由于 $A_1 \cup A_2 \cup \dots \cup A_k = U$ (见定义 1 (3)) 且 $A_i \subseteq A_i' (i=1, 2, \dots, k)$, 因此 $U \subseteq A_1' \cup A_2' \cup \dots \cup A_k'$, 各相关集涉及的数据是数据集 U 的扩展,或相关集类 A' 是划分 A 的扩展,这由条件 $A_i \subseteq A_i'$ 所决定,由此也可以看到 A 的相关集类并不唯一。针对实际问题时,相关集类的确定往往与数据的特性密切相关(见第 5 节的讨论)。

对于数据集 U 的划分 $A = \{A_1, A_2, \dots, A_k\}$, 设 $A' = \{A_1', A_2', \dots, A_k'\}$ 是 A 的相关集类,若每一粒 A_i 和其相关集 A_i'

($i=1,2,\dots,k$)被进一步分解,则得到细分的概念。

定义3 设 $A=\{A_1, A_2, \dots, A_k\}$ 是 U 的划分, $A'=\{A_1', A_2', \dots, A_k'\}$ 是其相关集类。若存在 U 的另一划分 $B=\{B_1, B_2, \dots, B_r\}$ 和其相关集类 $B'=\{B_1', B_2', \dots, B_r'\}$, 使对任意 $B_j \in B (j=1,2,\dots,r)$, 存在 $A_i \in A (1 \leq i \leq k)$, 满足 $B_j \subseteq A_i$, 则称 B 是 A 的细分; 同时, 若对于 B_j 的相关集 $B_j' \in B'$, 及 A_i 的相关集 $A_i' \in A'$, 也满足 $B_j' \subseteq A_i'$, 则称 B 是 A 的与相关集类关联的细分。

后面的讨论将涉及细分以及与相关集类关联的细分的性质, 现以结论的形式给出, 并予以证明。

结论 设 $C=\{C_1, C_2, \dots, C_s\}$, $B=\{B_1, B_2, \dots, B_r\}$ 和 $A=\{A_1, A_2, \dots, A_k\}$ 是 U 的划分, $C'=\{C_1', C_2', \dots, C_s'\}$, $B'=\{B_1', B_2', \dots, B_r'\}$ 和 $A'=\{A_1', A_2', \dots, A_k'\}$ 分别是它们的相关集类。则:

(1) 若 C 是 B 的细分, B 是 A 的细分, 则 C 是 A 的细分。

(2) 若 C 是 B 的与相关集类关联的细分, B 是 A 的与相关集类关联的细分, 则 C 是 A 的与相关集类关联的细分。

证明: (1) 对于任意 $C_t \in C (1 \leq t \leq s)$, 由于 C 是 B 的细分, 因此存在 $B_j \in B (1 \leq j \leq r)$, 使得 $C_t \subseteq B_j$ 。又因为 B 是 A 的细分, 所以存在 $A_i \in A (1 \leq i \leq k)$, 使得 $B_j \subseteq A_i$ 。因此由 $C_t \subseteq B_j$ 及 $B_j \subseteq A_i$, 可得 $C_t \subseteq A_i$ 。于是对于任意 $C_t \in C$, 存在 $A_i \in A$, 使得 $C_t \subseteq A_i$, 故 C 是 A 的细分。

(2) 当 C 是 B 的与相关集类关联的细分, 且 B 是 A 的与相关集类关联的细分时, 由定义3可知, C 必是 B 的细分, B 必是 A 的细分。由(1), C 是 A 的细分, 因此对于任意 $C_t \in C$, 存在 $A_i \in A$, 使得 $C_t \subseteq A_i$ 。考虑 C_t 的相关集 $C_t' (\in C')$ 和 A_i 的相关集 $A_i' (\in A')$, 可以证明 $C_t' \subseteq A_i'$ 。

事实上, 对于 $C_t \in C$, 因为 C 是 B 的与相关集类关联的细分, 所以存在 $B_j \in B$, 使得 $C_t \subseteq B_j$; 同时对于 C_t 的相关集 $C_t' (\in C')$ 和 B_j 的相关集 $B_j' (\in B')$, 满足 $C_t' \subseteq B_j'$ 。对于 $B_j \in B$, 因为 B 是 A 的与相关集类关联的细分, 所以存在 $A_h \in A$, 使得 $B_j \subseteq A_h$, 且对于 B_j 的相关集 $B_j' (\in B')$ 及 A_h 的相关集 $A_h' (\in A')$, 满足 $B_j' \subseteq A_h'$ 。所以由 $C_t' \subseteq B_j'$ 及 $B_j' \subseteq A_h'$, 有 $C_t' \subseteq A_h'$ 。再注意到 $C_t \subseteq B_j$ 并且 $B_j \subseteq A_h$, 因此 $C_t \subseteq A_h$ 。又因为上述已有 $C_t \subseteq A_i$, 所以 $A_h = A_i$ (因为 A_h 和 A_i 都是划分 A 中的粒, 且它们的交非空)。于是对于相关集 A_h' 和 A_i' 而言, 有 $A_h' = A_i'$ (因为 $A_i \subseteq A_h$, $A_h \subseteq A_i$ 且 $A_h = A_i$, 以及相关集类 A' 中不同相关集的交为空(见定义2)), 因此 $C_t' \subseteq A_i'$ 。

这样划分 C 不仅是划分 A 的细分, 即对于任意 $C_t \in C$, 存在 $A_i \in A$, 使得 $C_t \subseteq A_i$, 同时对于 C_t 的相关集 $C_t' (\in C')$ 和 A_i 的相关集 $A_i' (\in A')$, 也有 $C_t' \subseteq A_i'$, 故 C 是 A 的与相关集类关联的细分。

因此对数据集 U 而言, 其划分的逐步细分是粒被层层细化的过程, 粒涉及的相关集也随着粒的细分而被细分。

于是当 $B=\{B_1, B_2, \dots, B_r\}$ 是 $A=\{A_1, A_2, \dots, A_k\}$ 的细分时, 对于粒 $A_i \in A (i=1,2,\dots,k)$, 有 $A_i = B_{j_1} \cup B_{j_2} \cup \dots \cup B_{j_t}$, 其中 $B_{j_1}, B_{j_2}, \dots, B_{j_t} \in B (t \geq 1)$, 即粒 A_i 被分解为更精细的粒 $B_{j_1}, B_{j_2}, \dots, B_{j_t}$ 。此时划分 B 比划分 A 更精细。

当 $B=\{B_1, B_2, \dots, B_r\}$ 是 $A=\{A_1, A_2, \dots, A_k\}$ 的与相关集类关联的细分时, 对于 $B_j \in B$, 存在 $A_i \in A$, 使得 $B_j \subseteq A_i$, 同时相关集类 $B'=\{B_1', B_2', \dots, B_r'\}$ 与 $A'=\{A_1', A_2', \dots, A_k'\}$ 之间也具有与划分 B 和 A 之间类似的特性, 即当 $B_j' \in B'$ 时, 存在 $A_i' \in A'$, 这里 B_j' 和 A_i' 分别是 B_j 和 A_i 的相关集, 满足

$B_j' \subseteq A_i'$ 。此时 B' 比 A' 更精细。

对于数据集 U 的划分 $A=\{A_1, A_2, \dots, A_k\}$, 因为 $A_i \subseteq A_i (i=1,2,\dots,k)$, 所以 A 是自身的细分。同时 $A_i \not\subseteq A_i$ 说明 A 是自身的相关集类, 因此 A 也是自身的与相关集类关联的细分。

3 相关集类之间的数据关联

为讨论数据关联问题, 需要把不同的划分与其相关集类组合在一起, 形成由粒构成并涉及相关集类的数学结构。

3.1 关联于相关集类的粒化树

设 U 是数据集, 显然 $P_0=\{U\}$ 是 U 的划分。除此之外, 设 $P_1, P_2, \dots, P_n$ 是 U 的 n 个划分, $P_1', P_2', \dots, P_n'$ 分别是其相关集类, 其中 P_1 是 P_0 的细分, P_{i+1} 是 $P_i (i=1,2,\dots,n-1)$ 的与相关集类关联的细分。由结论(2)可知, 当 $j > i \geq 1$ 时, P_j 是 P_i 的与相关集类关联的细分, 此时 P_j 比 P_i, P_j' 比 P_i' 更精细。

令 $G=P_0 \cup P_1 \cup P_2 \cup \dots \cup P_n$, 则有 G 是 $P_0, P_1, P_2, \dots, P_n$ 中所有粒构成的集合。对于 $A_k \in G$, 存在 $i (0 \leq i \leq n)$, 使得 $A_k \in P_i$, 即 $A_k \subseteq U$ 。实际上包含关系“ $\subseteq$ ”是 G 上的偏序关系, 即满足自反性(对于 $A_k \in G$, 有 $A_k \subseteq A_k$)、反对称性(对于 $A_k, A_h \in G$, 如果 $A_k \subseteq A_h$ 且 $A_h \subseteq A_k$, 则 $A_k = A_h$)和传递性(对于 $A_q, A_k, A_h \in G$, 如果 $A_q \subseteq A_k$ 且 $A_k \subseteq A_h$, 则 $A_q \subseteq A_h$)。因此 G 和 $\subseteq$ 构成偏序集^[12], 记作 $T(U)=(G, \subseteq)$ 。偏序集 $T(U)=(G, \subseteq)$ 是一数学结构, 此时的 $T(U)$ 与相关集类 $P_1', P_2', \dots, P_n'$ 相联系。

定义4 (1) 结构 $T(U)=(G, \subseteq)$ 称为基于数据集 U 并与 n 个相关集类 $P_1', P_2', \dots, P_n'$ 关联的 n 层粒化树, 简称为粒化树。同时当不考虑相关集类 $P_1', P_2', \dots, P_n'$ 时, 结构 $T(U)=(G, \subseteq)$ 自身也称为 n 层粒化树或简称粒化树。

(2) 对于构成 $G=(P_0 \cup P_1 \cup P_2 \cup \dots \cup P_n)$ 的划分 $P_i (1 \leq i \leq k)$ 以及相关集类 P_i' , 称 P_i 和 P_i' 分别为 $T(U)$ 的第 i 层划分和第 i 层相关集类。

n 层粒化树 $T(U)=(G, \subseteq)$ 不仅是一种数学结构, 还关联于 n 个相关集类 $P_1', P_2', \dots, P_n'$ 。相关集类在文献[10, 11]的粒化树中未曾涉及, 是本文数据关联描述的前提。 U 在 $T(U)$ 中的出现表明粒化树基于 U 之上, “ n 层”指明了 $T(U)=(G, \subseteq)$ 的层数, “树”意味着 $T(U)=(G, \subseteq)$ 在平面上的树型结构。

例如, 设数据集 $U=\{a_1, a_2, a_3, a_4, a_5, a_6, a_7, a_8, a_9\}$ 。令 $G=P_0 \cup P_1 \cup P_2 \cup P_3 (P_0=\{U\})$, 其中 $P_1=\{\{a_1, a_2, a_3, a_4, a_5\}, \{a_6, a_7, a_8, a_9\}\}$, $P_2=\{\{a_1, a_2, a_3\}, \{a_4, a_5\}, \{a_6, a_7, a_8, a_9\}\}$, $P_3=\{\{a_1, a_2\}, \{a_3\}, \{a_4, a_5\}, \{a_6, a_7\}, \{a_8, a_9\}\}$ 。则 P_0, P_1, P_2, P_3 都是 U 的划分, P_{i+1} 是 $P_i (i=0,1,2)$ 的细分。此时 $T(U)=(G, \subseteq)$ 是 3 层粒化树。若把 $T(U)=(G, \subseteq)$ 的各层画在平面上(见图1), 其形状是图论^[14]中的树。

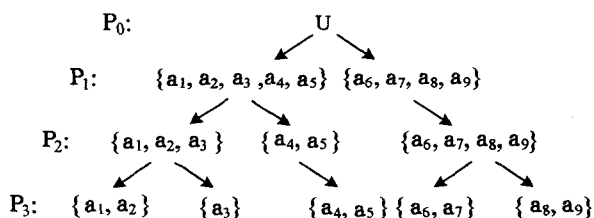


图1 3层粒化树 $T(U)=(G, \subseteq)$

该图中, U 是树根, 对于粒 $A_k, A_h \in G$, 从 A_k 到 A_h 存在有向线段(即箭头), 当且仅当 $A_k \in P_i$ 且 $A_h \in P_{i+1}$ ($i=0, 1, 2$), 满足 $A_h \subseteq A_k$. P_i ($i=0, 1, 2, 3$) 是 $T(U)=(G, \subseteq)$ 的第 i 层划分. P_3 中的粒 $\{a_1, a_2\}, \{a_3\}, \{a_4, a_5\}, \{a_6, a_7\}, \{a_8, a_9\}$ 是树叶. 注意同一粒可位于不同的层, 如 $\{a_4, a_5\} \in P_2$ 且 $\{a_4, a_5\} \in P_3$.

3.2 数据关联的定义

如果把图 1 中的 3 层粒化树记作 $T(U)_1=(G_1, \subseteq)$, 这里 $G_1=P_0 \cup P_1 \cup P_2 \cup P_3$, 则可构造不同的粒化树. 比如设 Q_0, Q_1, Q_2, Q_3, Q_4 是 $U=\{a_1, a_2, a_3, a_4, a_5, a_6, a_7, a_8, a_9\}$ 的划分, 其中 $Q_0=\{U\}, Q_1=\{\{a_1, a_2, a_3\}, \{a_4, a_5, a_6\}, \{a_7, a_8, a_9\}\}, Q_2=\{\{a_1, a_2\}, \{a_3\}, \{a_4, a_5, a_6\}, \{a_7, a_8\}, \{a_9\}\}, Q_3=\{\{a_1, a_2\}, \{a_3\}, \{a_4, a_5\}, \{a_6\}, \{a_7\}, \{a_8\}, \{a_9\}\}, Q_4=\{\{a_1\}, \{a_2\}, \{a_3\}, \{a_4\}, \{a_5\}, \{a_6\}, \{a_7\}, \{a_8\}, \{a_9\}\}$, 显然 Q_1, Q_2, Q_3, Q_4 是不同于 P_1, P_2, P_3 的划分, 且 Q_{i+1} 是 Q_i ($i=0, 1, 2, 3$) 的细分. 令 $G_2=Q_0 \cup Q_1 \cup Q_2 \cup Q_3 \cup Q_4$, 则得到不同于 $T(U)_1$ 的 4 层粒化树 $T(U)_2=(G_2, \subseteq)$. 同时可以对 $T(U)_1$ 和 $T(U)_2$ 每一层划分中的粒指定相关集, 得到每一层划分的相关集类, 使得 $T(U)_1$ 和 $T(U)_2$ 成为与相关集类关联的粒化树.

一般地, 设 $T(U)_1=(G_1, \subseteq)$ 是基于数据集 U 并与 m 个相关集类 $P_1', P_2', \dots, P_m'$ 关联的 m 层粒化树, 其中 $G_1=P_0 \cup P_1 \cup \dots \cup P_m$ ($P_0=\{U\}$), 且 P_{i+1} 是 P_i ($i=1, 2, \dots, m-1$) 的与相关集类关联的细分; 同时设 $T(U)_2=(G_2, \subseteq)$ 是基于数据集 U 并与 n 个相关集类 $Q_1', Q_2', \dots, Q_n'$ 关联的 n 层粒化树, 其中 $G_2=Q_0 \cup Q_1 \cup \dots \cup Q_n$ ($Q_0=\{U\}$), 且 Q_{j+1} 是 Q_j ($j=1, 2, \dots, n-1$) 的与相关集类关联的细分.

对于数据 $x \in U$, 考虑粒化树 $T(U)_1=(G_1, \subseteq)$ 的第 i 层划分 P_i ($0 \leq i \leq m$), 由于 P_i 是 U 的划分, 因此一定存在粒 $A_k \in P_i$, 使得 $x \in A_k$. 再考虑 $T(U)_2=(G_2, \subseteq)$ 的第 j 层划分 Q_j ($0 \leq j \leq n$), 此时存在 $B_h \in Q_j$, 使得 $x \in B_h$. 因此 $x \in A_k \cap B_h$.

定义 5 (1) 对于数据 $x \in U$, 如果 $x \in A_k \cap B_h$, 则称 x 是 (A_k, B_h) -关联数据, 其中 $A_k \in P_i$ 及 $B_h \in Q_j$, 这里 P_i 是 $T(U)_1=(G_1, \subseteq)$ 的第 i 层划分, Q_j 是 $T(U)_2=(G_2, \subseteq)$ 的第 j 层划分.

(2) 设 $x \in U$ 且 x 是 (A_k, B_h) -关联数据, 其中 $A_k \in P_i$ 及 $B_h \in Q_j$, 对于 A_k 的相关集 $A_k' \in P_i'$ 及 B_h 的相关集 $B_h' \in Q_j'$, 当 $y \in A_k'$ 及 $z \in B_h'$ 时, 称数据 y 和 z 是 (x, i, j) -关联的.

由定义 5(2) 可知, 若 y 和 z 是 (x, i, j) -关联的, 则 $x \in U$, y 和 z 分别属于相应的相关集, y 或 z 可以不是 U 中的数据.

至此通过与相关集类关联的粒化树 $T(U)_1$ 和 $T(U)_2$, 对 y 和 z 的 (x, i, j) -关联进行了定义. 对于数据 x, y 和 z 而言, x 是 (A_k, B_h) -关联数据, 即 $x \in U$ 且 $x \in A_k \cap B_h$, 其中 A_k 是 $T(U)_1$ 第 i 层划分 P_i 中的粒, B_h 是 $T(U)_2$ 第 j 层划分 Q_j 中的粒, 同时 y 是 A_k 的相关集 A_k' 中的数据, z 是 B_h 的相关集 B_h' 中的数据. 此时 U 中的数据 x 把 A_k 相关集 A_k' 中的数据 y 与 B_h 相关集 B_h' 中的数据 z 关联在了一起.

三元组 (x, i, j) 中的数值 i 和 j 分别表示 $T(U)_1$ 和 $T(U)_2$ 中划分 P_i 和 Q_j 的层次, 也可认为是粒 $A_k (\in P_i)$ 和 $B_h (\in P_j)$ 以及相关集 $A_k' (\in P_i')$ 和 $B_h' (\in Q_j')$ 的层次. 由于粒和相关集均由数据构成, 因此 i 实际是 A_k 和 A_k' 中数据的数值表示, j 是 B_h 和 B_h' 中数据的数值表示. 数值 i 或 j 的变化将影响 y 和 z 的 (x, i, j) -关联程度, 是关联紧密性的数字化体现.

第 5 节将讨论数值 i 和 j 与数据 y 和 z 的 (x, i, j) -关联之间的联系.

当讨论不强调数据以及划分的层次时, 将把 y 和 z 的 (x, i, j) -关联简称为数据关联.

4 数据关联的研究

当数据 y 和 z 是 (x, i, j) -关联时, 分析数据 x, y, z 之间的关系, 讨论数值 i 或 j 的变化对关联程度的影响, 研究数据关联的性质等是需要面对的内容. 但考虑篇幅问题, 本节仅给出基本的结论, 相关性质在第 5 节通过实例讨论说明.

设 $T(U)_1=(G_1, \subseteq)$ 是基于数据集 U 并与 m 个相关集类 $P_1', P_2', \dots, P_m'$ 关联的 m 层粒化树, $G_1=P_0 \cup P_1 \cup \dots \cup P_m$ ($P_0=\{U\}$), 且 P_{i+1} 是 P_i ($i=1, 2, \dots, m-1$) 的与相关集类关联的细分. $T(U)_2=(G_2, \subseteq)$ 是基于 U 并与 n 个相关集类 $Q_1', Q_2', \dots, Q_n'$ 关联的 n 层粒化树, $G_2=Q_0 \cup Q_1 \cup \dots \cup Q_n$ ($Q_0=\{U\}$), 且 Q_{j+1} 是 Q_j ($j=1, 2, \dots, n-1$) 的与相关集类关联的细分. 本节的讨论中, $T(U)_1$ 和 $T(U)_2$ 均如此约定.

考虑 $T(U)_1$ 的第 i 层划分 P_i 和相关集类 P_i' ($1 \leq i \leq m$), 以及 $T(U)_2$ 的第 j 层划分 Q_j 和相关集类 Q_j' ($1 \leq j \leq n$), 设 $A \in P_i$ 且 $C \in P_i'$, 以及 $B \in Q_j$ 且 $D \in Q_j'$, 则有如下结论成立.

定理 1 对于 $x \in U$, 如果 $x \in A$ 且 $x \in C$, 则 A 的相关集是 C . 如果 $x \in B$ 且 $x \in D$, 则 B 的相关集是 D .

证明: 设 A 的相关集为 $A' \in P_i'$, 则由粒和其相关集的关系(见定义 2)可知 $A \subseteq A'$. 因为 $x \in A$, 因此 $x \in A'$. 又因为 $x \in C$, 因此 $A' \cap C \neq \emptyset$. 由于同一相关集类中不同相关集的交为空(见定义 2), 因此 $A'=C$, 即 A 的相关集是 C . 同理可证 B 的相关集是 D .

给出定理 1 是为了证明定理 2, 从而确定数据 y 和 z 是 (x, i, j) -关联的充分必要条件.

定理 2 数据 y 和 z 是 (x, i, j) -关联的充分必要条件是 $x \in U$, 并且 $x, y \in A_k'$ 及 $x, z \in B_h'$, 这里 $A_k' \in P_i'$ 且 $B_h' \in Q_j'$.

证明: 设 y 和 z 是 (x, i, j) -关联的. 则由定义 5(2) 可知, $x \in U$ 且 x 是 (A_k, B_h) -关联数据, 即 $x \in A_k \cap B_h$, 其中 $A_k \in P_i$ 且 $B_h \in Q_j$; 同时 $y \in A_k'$ 及 $z \in B_h'$, 这里 $A_k' \in P_i'$ 且 A_k' 是 A_k 的相关集, $B_h' \in Q_j'$ 且 B_h' 是 B_h 的相关集. 由粒与其相关集的关系可知 $A_k \subseteq A_k'$ 且 $B_h \subseteq B_h'$ (见定义 2), 因此由 $x \in A_k \cap B_h$, 有 $x \in A_k'$ 及 $x \in B_h'$. 故 $x, y \in A_k'$ 及 $x, z \in B_h'$.

反之, 设 $x \in U$, 并且 $x, y \in A_k'$ 及 $x, z \in B_h'$, 其中 $A_k' \in P_i'$ 且 $B_h' \in Q_j'$. 因为 $x \in U$, 以及 P_i 和 Q_j 都是 U 的划分, 所以存在粒 $A_k \in P_i$ 及粒 $B_h \in Q_j$, 使得 $x \in A_k$ 且 $x \in B_h$, 即 $x \in A_k \cap B_h$, 所以 x 是 (A_k, B_h) -关联数据. 再注意到 $x \in A_k$ 且 $x \in A_k'$, 以及 $x \in B_h$ 且 $x \in B_h'$, 因此由定理 1 可知, A_k 的相关集是 A_k' , B_h 的相关集是 B_h' . 由于 $y \in A_k'$ 及 $z \in B_h'$, 由定义 5(2) 可知, 数据 y 和 z 是 (x, i, j) -关联的.

因此条件 $x, y \in A_k'$ 及 $x, z \in B_h'$ 刻画了数据 y 和 z 是 (x, i, j) -关联的含义. 由此可知, 数据集 U 中的数据 x 作为相关集 A_k' 和 B_h' 中的公共数据, 把 A_k' 的数据 y 和 B_h' 的数据 z 关联到一起. 实际上, 第 1 节(1)-(5)中数据关联的例子都可归结为 y 和 z 是 (x, i, j) -关联的形式, 所以定义 5(2) 给出了实际中数据关联的描述方法, i 和 j 是相关信息的数值表示.

基于定理 2 可对数据关联的性质展开研究, 但因篇幅所限, 第 5 节将通过实例对其进行讨论. 这将涉及 (x, i, j) 中数值 i 和 j 的变化或比较, 因此给出相关的约定如下:

(1) $(q, r) = (i, j)$ 当且仅当 $q = i$ 且 $r = j$ (i, j, q, r 是自然数);

(2) $(q, r) > (i, j)$ 当且仅当 $q > i$ 且 $r \geq j$, 或者 $q \geq i$ 且 $r > j$;

(3) $(q, r) \geq (i, j)$ 当且仅当 $(q, r) > (i, j)$ 或者 $(q, r) = (i, j)$ 。

第5节将围绕如下问题对例子进行讨论:

(1) 如果 y 和 z 既是 (x_1, i, j) -关联的, 又是 (x_2, i, j) -关联的, 那么 x_1 和 x_2 具有怎样的联系?

(2) 如果 y_1 和 z_1 是 (x, i, j) -关联的, 同时 y_2 和 z_2 也是 (x, i, j) -关联的, 那么 y_1 和 y_2 以及 z_1 和 z_2 具有怎样的联系?

(3) 如果 y 和 z 是 (x, q, r) -关联的, 那么当 $(q, r) > (i, j)$ 时, y 和 z 是否一定是 (x, i, j) -关联的?

(4) 如果 y 和 z 既是 (x, q, r) -关联的, 又是 (x, i, j) -关联的, 那么 (q, r) 和 (i, j) 的比较是否可以体现关联的紧密程度?

5 基于实例的数据关联讨论

本节给出一实例, 涉及的讨论将围绕上述几个问题。由于粒化树是这些问题的结构支撑, 因此下文将从同一数据集上不同粒化树的构造开始, 并使每一粒对应确定的相关集。

例3 设数据集 U 表示某大学计算机学院全体学生的集合, 即 $U = \{x \mid x \text{ 是某大学计算机学院的学生}\}$ 。

基于 U 构造4层粒化树 $T(U)_1 = (G_1, \subseteq)$, 具体过程如下。

令 $P_1 = \{A \mid A \subseteq U \text{ 且 } A \text{ 中的学生来自同一县市}\}$, 则 P_1 是 U 的划分, 由来自同一县市的学生确定。当 $A \in P_1$ 时, 令 A' 由 A 中学生涉及县市的所有人员构成, 即 A' 包含该县市人员的全体, 包括到计算机学院读书的学生, 于是 $A \subseteq A'$ 。指定 A' 为 A 的相关集, 令 $P_1' = \{A' \mid A \in P_1\}$ 。由于当 $A_i, A_j \in P_1$ 且 $A_i \neq A_j$ 时, 有 $A_i' \cap A_j' = \emptyset$ (即不同县市人员的集合交为空), 因此 P_1' 是 P_1 的相关集类。

令 $P_2 = \{A \mid A \subseteq U \text{ 且 } A \text{ 中的学生来自同一乡镇}\}$, 这意味着 P_2 是将 P_1 中每一县市学生构成的粒分解为若干个乡镇学生构成的粒的结果, 显然 P_2 是 P_1 的细分。同时对于 $A \in P_2$, 令 A' 由 A 中学生涉及乡镇的所有人员 (包括就读于计算机学院的学生) 构成, 则 $A \subseteq A'$ 。指定 A' 为 A 的相关集。令 $P_2' = \{A' \mid A \in P_2\}$, 显然当 $A_i, A_j \in P_2$ 且 $A_i \neq A_j$ 时, 有 $A_i' \cap A_j' = \emptyset$, 于是 P_2' 构成 P_2 的相关集类。注意到当 $A_i \in P_2, A_i \in P_1$ 且 $A_i \subseteq A_k$ 时, 必有 $A_i' \subseteq A_k'$, 这里 $A_i' \in P_2'$ 且 $A_k' \in P_1'$, A_i' 和 A_k' 分别是 A_i 和 A_k 的相关集。所以 P_2 是 P_1 的与相关集类关联的细分。

令 $P_3 = \{A \mid A \subseteq U \text{ 且 } A \text{ 中的学生来自同一村庄}\}$, 即 P_3 是 P_2 中每一乡镇分解为村庄的结果, 则 P_3 是 P_2 的细分。对于 $A \in P_3$, 令 A' 由 A 涉及村庄的所有人员构成, 指定 A' 为 A 的相关集, 显然 $A \subseteq A'$ 。令 $P_3' = \{A' \mid A \in P_3\}$, 则 P_3' 构成 P_3 的相关集类, 此时 P_3 是 P_2 的与相关集类关联的细分。

令 $P_4 = \{A \mid A \subseteq U \text{ 且 } A \text{ 中的学生来自同一村庄并且性别相同}\}$, 即 P_4 是将 P_3 中的每一粒按性别细分的结果。对于 $A \in P_4$, 令 A' 由 A 涉及村庄的所有男性或女性人员构成, 此时 $A \subseteq A'$ 。指定 A' 为 A 的相关集, 于是 $P_4' = \{A' \mid A \in P_4\}$ 形成 P_4 的相关集类, P_4 是 P_3 的与相关集类关联的细分。

上述构造表明, 对于划分 P_i 和 P_j 以及相关集类 P_i' 和 P_j' ($1 \leq i, j \leq 4$), 当 $j > i$ 时, 如果 $A_i \in P_j$, 则存在 $A_k \in P_i$, 使得 $A_i \subseteq A_k$ 。此时粒 A_i 和 A_k 的相关集 $A_i' (\in P_j')$ 及 $A_k' (\in$

$P_i')$ 也满足 $A_i' \subseteq A_k'$ 。同时对于 $P_i' (i=1, 2, 3, 4)$ 中的相关集 $A_k', A_k' \in P_i'$, 若 A_k' 与 A_k' 不同, 则 $A_k' \cap A_k' = \emptyset$ 。因此当 $j > i \geq 1$ 时, P_j 是 P_i 的与相关集类关联的细分。

令 $G_1 = P_0 \cup P_1 \cup P_2 \cup P_3 \cup P_4$ ($P_0 = \{U\}$), 则 $T(U)_1 = (G_1, \subseteq)$ 是基于 U 并与4个相关集类 P_1', P_2', P_3', P_4' 关联的4层粒化树, 且 P_{i+1} 是 $P_i (i=1, 2, 3)$ 的与相关集类关联的细分。

基于 U 还可构造5层粒化树 $T(U)_2 = (G_2, \subseteq)$, 具体过程如下。

令 $Q_1 = \{B_1, B_2, B_3\}$, 其中 B_1 由计算机学院的理科生构成, B_2 由工科生构成, B_3 由教育类学生构成, 显然 $B_i \subseteq U (i=1, 2, 3)$ 。同时 U 就包含理科、工科和教育3类学生, 于是 Q_1 构成 U 的划分。令 B_i' 由该大学 (包含计算机学院) 的理科生构成, B_2' 由该大学的工科生构成, B_3' 由该大学的教育类学生构成, 显然 $B_i \subseteq B_i' (i=1, 2, 3)$ 。指定 B_i' 为 $B_i (i=1, 2, 3)$ 的相关集, 令 $Q_1' = \{B_1', B_2', B_3'\}$, 显然 B_1', B_2' 和 B_3' 之间互不相交, 所以 Q_1' 构成了 Q_1 的相关集类。

Q_2 是 $Q_1 = \{B_1, B_2, B_3\}$ 的与相关集类关联的细分, 是通过年级对 Q_1 中每一粒分解的结果。对于 $B_i \in Q_1 (i=1, 2, 3)$, B_i 被分为一年级、二年级、三年级和四年级这4个粒, Q_2 由所有这样细分的粒构成, 即若 $B_h \in Q_2$, 则 B_h 是计算机学院同学科且同年级学生构成的粒。相应地, $Q_1' = \{B_1', B_2', B_3'\}$ 中的 $B_i' (i=1, 2, 3)$ 被细分为一年级到四年级学生的子集, 从而得到 Q_2' 。对于 $B_h' \in Q_2', B_h'$ 由该大学 (不仅计算机学院) 同学科且同年级的学生构成, 指定 B_h' 为 B_h 的相关集, 显然 $B_h \subseteq B_h'$ 。由于当 $B_i', B_j' \in Q_2'$ 且 B_i' 与 B_j' 不同时, 有 $B_i' \cap B_j' = \emptyset$, 因此 Q_2' 构成了 Q_2 的相关集类。

Q_3 是 Q_2 的与相关集类关联的细分, Q_3 中的粒是 Q_2 中每一粒以“学生英语六级通过与否”细分的结果。同时对 Q_2 的相关集类 Q_2' 中的每一相关集也以学生英语六级通过与否进行分解, 从而得到 Q_3 的相关集类 Q_3' 。

Q_4 是 Q_3 的与相关集类关联的细分, 通过“数学成绩优秀与否”对 Q_3 中每一粒细分后得到 Q_4 。同时 Q_3 相关集类 Q_3' 中每一相关集也以学生的数学成绩优秀与否进行分解, 从而得到 Q_4 的相关集类 Q_4' 。

Q_5 是 Q_4 的与相关集类关联的细分, 通过学生的性别对 Q_4 中每一粒细分后得到 Q_5 。相应地, Q_4 相关集类 Q_4' 中的每一相关集也以学生性别进行分解, 从而得到 Q_5 的相关集类 Q_5' 。

显然划分 Q_1, Q_2, Q_3, Q_4, Q_5 和相关集类 $Q_1', Q_2', Q_3', Q_4', Q_5'$ 与上述的划分 P_1, P_2, P_3, P_4 及相关集类 P_1', P_2', P_3', P_4' 不同, 且当 $t > s \geq 1$ 时, Q_t 是 Q_s 的与相关集类关联的细分。

令 $G_2 = Q_0 \cup Q_1 \cup Q_2 \cup Q_3 \cup Q_4 \cup Q_5$ ($Q_0 = \{U\}$), 则 $T(U)_2 = (G_2, \subseteq)$ 是基于数据集 U , 并与5个相关集类 $Q_1', Q_2', Q_3', Q_4', Q_5'$ 关联的5层粒化树, 且 Q_{s+1} 是 $Q_s (s=1, 2, 3, 4)$ 的与相关集类关联的细分。

这样通过 U 中数据 (大学生) 满足的性质构造了4层和5层粒化树 $T(U)_1 = (G_1, \subseteq)$ 和 $T(U)_2 = (G_2, \subseteq)$ 。由此可围绕第4节后端的问题展开讨论, 形成如下几方面的研究。

(1) 对于 $x_1, x_2 \in U$, 设 y 和 z 是 (x_1, i, j) -关联的, 同时又是 (x_2, i, j) -关联的 ($0 \leq i \leq 4$ 且 $0 \leq j \leq 5$)。由定理2可知, $x_1, y \in A_k'$ 及 $x_1, z \in B_h'$, 其中 $A_k' \in P_i'$ 且 $B_h' \in Q_j'$; 同时 $x_2,$

$y \in A_i'$ 及 $x_2, z \in B_j'$, 其中 $A_i' \in P_i'$ 且 $B_j' \in Q_j'$ 。由 $y \in A_k'$ 和 $y \in A_i'$ 及相关集类 P_i' 中不同相关集的交为空可知 $A_k' = A_i'$ 。同理 $B_h' = B_j'$ 。因此 $x_1, x_2 \in A_k' (= A_i')$ 及 $x_1, x_2 \in B_h' (= B_j')$ 。这说明数据 x_1 和 x_2 具有粒 A_k' 及粒 B_h' 蕴含的性质, 此时可称 x_1 和 x_2 既是 A_k' -等价的, 也是 B_h' -等价的。因此由 y 和 z 的 (x_1, i, j) -关联以及 (x_2, i, j) -关联可推得 x_1 和 x_2 的 A_k' -等价和 B_h' -等价。

(2) 设 y_1 和 z_1 是 (x, i, j) -关联的, 同时 y_2 和 z_2 也是 (x, i, j) -关联的 ($0 \leq i \leq 4$ 且 $0 \leq j \leq 5$)。则由定理 2 可知, $x, y_1 \in A_k'$ 及 $x, z_1 \in B_h'$, 其中 $A_k' \in P_i'$ 且 $B_h' \in Q_j'$; 同时 $x, y_2 \in A_i'$ 及 $x, z_2 \in B_j'$, 其中 $A_i' \in P_i'$ 且 $B_j' \in Q_j'$ 。由于相关集类 P_i' 或 Q_j' 中不同相关集的交为空, 因此由 $x \in A_k'$ 且 $x \in A_i'$, 以及 $x \in B_h'$ 且 $x \in B_j'$, 得 $A_k' = A_i'$ 及 $B_h' = B_j'$ 。因此 $y_1, y_2 \in A_k' (= A_i')$, 并且 $z_1, z_2 \in B_h' (= B_j')$, 这说明 y_1 和 y_2 是 A_k' -等价的, z_1 和 z_2 是 B_h' -等价的。因此当 y_1 和 z_1 是 (x, i, j) -关联及 y_2 和 z_2 是 (x, i, j) -关联时, 可推出 y_1 和 y_2 是 A_k' -等价的, 且 z_1 和 z_2 是 B_h' -等价的结论。

(3) 若 y 和 z 是 (x, q, r) -关联的, 则当 $(q, r) > (i, j)$ 时, y 和 z 必是 (x, i, j) -关联的 ($0 \leq q, i \leq 4$ 且 $0 \leq r, j \leq 5$)。事实上, 当 y 和 z 是 (x, q, r) -关联时, 由定理 2 可知, $x, y \in A_k'$ 及 $x, z \in B_h'$, 其中 $A_k' \in P_q$ 且 $B_h' \in Q_r$ 。设 $A_k \in P_q$ 且 A_k' 是 A_k 的相关集, $B_h \in Q_r$ 且 B_h' 是 B_h 的相关集。当 $(q, r) > (i, j)$ 时, 有 $q > i$ 且 $r \geq j$, 或 $q \geq i$ 且 $r > j$ 。此时 P_q 是 P_i 的与相关集类关联的细分, Q_r 是 Q_j 的与相关集类关联的细分, 因此存在粒 $A_i \in P_i$ 及 $B_j \in Q_j$, 使得 $A_k \subseteq A_i$ 且 $B_h \subseteq B_j$, 同时对于 A_i 的相关集 $A_i' \in P_i'$ 及 B_j 的相关集 $B_j' \in Q_j'$, 有 $A_k' \subseteq A_i'$ 并且 $B_h' \subseteq B_j'$ (见定义 3(2))。因此 $x, y \in A_i'$ 及 $x, z \in B_j'$ 。仍由定理 2 可知, y 和 z 是 (x, i, j) -关联的。

(4) 如果 y 和 z 既是 (x, q, r) -关联的 ($0 \leq q \leq 4$ 且 $0 \leq r \leq 5$), 又是 (x, i, j) -关联的 ($0 \leq i \leq 4$ 且 $0 \leq j \leq 5$), 那么当 $(q, r) > (i, j)$ 时, 可对关联的紧密程度进行分析: 由定理 2 可知, y 和 z 的 (x, q, r) -关联等价于 $x, y \in A_k'$ 以及 $x, z \in B_h'$, 其中 $A_k' \in P_q$ 且 $B_h' \in Q_r$ 。同样, y 和 z 的 (x, i, j) -关联等价于 $x, y \in A_i'$ 及 $x, z \in B_j'$, 其中 $A_i' \in P_i'$ 且 $B_j' \in Q_j'$ 。由于 $(q, r) > (i, j)$, 因此 P_q 是 P_i 的与相关集类关联的细分 (包括 $P_i = P_q$ 的情况), Q_r 是 Q_j 的与相关集类关联的细分 (包括 $Q_j = Q_r$ 的情况)。由于 y 和 z 是 (x, q, r) -关联的, 因此 x 是 (A_k, B_h) -关联数据, 即 $x \in A_k \cap B_h$, 其中 $A_k \in P_q$ 及 $B_h \in Q_r$, 由定理 1 可知, 上述 A_k' 和 B_h' 分别是 A_k 和 B_h 的相关集。同样因为 y 和 z 是 (x, i, j) -关联的, 所以 x 是 (A_i, B_j) -关联数据, 即 $x \in A_i \cap B_j$, 这里 $A_i \in P_i$ 及 $B_j \in Q_j$, 且 A_i' 和 B_j' 分别是 A_i 和 B_j 的相关集。由 $x \in A_k$ 且 $A_k \in P_q, x \in A_i$ 且 $A_i \in P_i$, 以及 P_q 是 P_i 的与相关集类关联的细分, 有 $A_k \subseteq A_i$, 于是 $A_k' \subseteq A_i'$ 。同理 $B_h' \subseteq B_j'$ 。这样 $x, y \in A_k'$ 且 $x, y \in A_i'$ 以及 $A_k' \subseteq A_i'$ 可解释为 x 和 y 的 A_k' -等价 (见(1)) 比 A_i' -等价更紧密。同样 $x, z \in B_h'$ 且 $x, z \in B_j'$ 以及 $B_h' \subseteq B_j'$ 意味着 x 和 z 的 B_h' -等同比 B_j' -等价更紧密。对这些通过 $(q, r) > (i, j)$ 的数值比较予以了确定, 这是数据信息的数字化体现。

上述讨论是对第 4 节后端问题的回答, 由此可以看到定理 2 的支撑作用。如下给出一些更具体的分析。

(5) 设 $x \in U$, 考虑 $T(U)_1$ 的第 3 层划分 P_3 和相关集类 P_3' , 以及 $T(U)_2$ 的第 4 层划分 Q_4 和相关集类 Q_4' 。由于 P_3 和 Q_4 都是 U 的划分, 因此存在粒 $A_k \in P_3$ 且 $B_h \in Q_4$, 使得 $x \in A_k$ 及 $x \in B_h$, 即 $x \in A_k \cap B_h$, 此时 x 是 (A_k, B_h) -关联数

据。因此对于 $y \in A_k' (= P_3')$ 及 $z \in B_h' (= Q_4')$, y 和 z 是 $(x, 3, 4)$ -关联的 (见定义 5(2))。实际上, 由 P_3' 的构造可知, $A_k' (= P_3')$ 表示某村庄的全体人员, 数值 3 是村庄人员的数字化表示。同样由 Q_4' 的构造可知, $B_h' (= Q_4')$ 由同学科、同年级、英语六级通过 (或未通过)、数学成绩优秀 (或不优秀) 的大学生构成, 数值 4 是这些特性的数字化信息。数据 y 和 z 的 $(x, 3, 4)$ -关联展示了来自村庄的大学生 x 将该村庄的人员 y 与其具有共同特性的大学同学 z 关联在一起的事实。

(6) 因为 $(3, 4) > (2, 3)$, 所以如果 y 和 z 是 $(x, 3, 4)$ -关联的, 则有 y 和 z 必是 $(x, 2, 3)$ -关联的 (见(3))。事实上, 如果某村庄人员的集合记作 A_k' (即 P_3' 的相关集), 包含该村庄乡镇人员的集合记作 A_i' (即 P_2' 的相关集), 则 $A_k' \subseteq A_i'$ 。同时如果同学科、同年级、同英语和同数学成绩的大学生的集合记作 B_h' (即 Q_4' 的相关集), 且同学科、同年级和同英语成绩的大学生的集合记作 B_j' (即 Q_3' 的相关集), 则有 $B_h' \subseteq B_j'$ 。此时 y 和 z 的 $(x, 3, 4)$ -关联是 A_k' 与 B_h' 之间的数据关联, 由定理 2 可知, $x, y \in A_k'$ 且 $x, z \in B_h'$ 。由于 $A_k' \subseteq A_i'$ 且 $B_h' \subseteq B_j'$, 因此 $x, y \in A_i'$ 且 $x, z \in B_j'$, 于是 y 和 z 是 $(x, 2, 3)$ -关联的。直观的解释就是, 若大学生能建立起村庄人员与大学生的关联, 则必然能建立起乡镇人员与更大范围大学生的关联。由于村庄与大学生的关联显然比乡镇与大学生的关联更紧密, 因此可认为 y 和 z 的 $(x, 3, 4)$ -关联比 $(x, 2, 3)$ -关联更紧密, 并通过 $(3, 4) > (2, 3)$ 得以数字化的体现。实际上, $x, y \in A_k'$ 且 $x, y \in A_i'$, 以及 $A_k' \subseteq A_i'$ 表明 x 和 y 的 A_k' -等同比 A_i' -等价更紧密 (见(4)), 同时 $x, z \in B_h'$ 且 $x, z \in B_j'$ 以及 $B_h' \subseteq B_j'$ 表明 x 和 z 的 B_h' -等同比 B_j' -等价更紧密, 这些都决定了数据关联的更紧密性。所以关联的更紧密性由更紧密的数据等价所决定, 并通过数值比较予以了体现。

(7) 上述粒化树的构造与数据关联的讨论建立了具体问题的数学模型, 为实际问题的算法模拟和程序编制提供了支撑, 是数据信息自动管理和网络查询的基础。

结束语 数据关联的定义, 展示了公共数据连接两类数据的形式化方法。而针对数据关联的讨论, 产生了与形式化表示互为等价的条件, 由此可推得与数据关联密切相关的性质。这些都是本文所做的工作, 但是由于篇幅的限制, 性质的讨论仅通过第 5 节的实例进行了解释性的分析。实际上, 如果仍把定理 2 作为研究的支撑, 那么通过对诸如数据等同、关联紧密、等同紧密等概念的定义, 可使数据关联的研究更为正式或规范, 这可作为今后深入讨论的内容。

如果总结工作中的特点, 那么相关集以及相关集类的引入体现了研究的特色。相关集的确定源于对数据集划分后, 与粒密切相关的一类数据, 同时通过不同性质对粒的逐步细化以及对相关集地层层细分, 使得与相关集类关联的不同粒化树得以产生, 由此形成了数据关联形式化表示的支撑环境。同一数据集上不同粒化树的构建有别于文献[10]中关联链基于不同数据集的做法, 也与文献[11]中数据关联涉及两个数据集的方法存在区别, 这是作者求变思想的体现。

参考文献

- [1] ZHANG Jun-bo, LI Tian-rui, RUAN Da. Neighborhood Rough Sets for Dynamic Data Mining[J]. International Journal of Intelligent Systems, 2012, 27(4): 317-342.

实验结果表明,本文的压缩融合方法有利于提高图像融合质量,具有较好的实用性。随着图像数量的增加,更显示出本算法的优势,即减小了计算量,降低了对传感器硬件的要求。

参 考 文 献

- [1] PETROVIĆ V, DIMITRIJEVIĆ V. Focused pooling for image fusion evaluation[J]. *Information Fusion*, 2015, 22: 119-126.
 - [2] DONOHO D L. Compressed sensing [J]. *Information Theory, IEEE Transactions on Information Theory*, 2006, 52(4): 1289-1306.
 - [3] WAN J, ZHOU S. Features extraction based on wavelet packet transform for B[C]// 2010 3rd International Congress on Image and Signal Processing (CISP). IEEE, 2010: 949-955.
 - [4] ZHANG Yi-zhuo, MA Lin, XU Lei, et al. Wood board texture classification based on genetic fusion of wavelet and curvelet features[J]. *Journal of Beijing Forestry University*, 2014, 36(2): 119-124. (in Chinese)
张怡卓, 马琳, 许雷, 等. 基于小波与曲线遗传融合的木材纹理分类[J]. *北京林业大学学报*, 2014, 36(2): 119-124.
 - [5] BAI X. Infrared and visual image fusion through feature extraction by morphological sequential toggle operator[J]. *Infrared Physics & Technology*, 2015, 71: 77-86.
 - [6] LI Chao, ZHANG Yi-zhuo, YU Hui-ling, et al. Dualtree complex wavelet feature fusion and wood board collaborative detection by compressed Sensing[J]. *Electric Machines and Control*, 2015, 19(8): 81-87. (in Chinese)
李超, 张怡卓, 于慧伶, 等. 双树复小波特征融合的板材压缩感知协同检测与分选[J]. *电机与控制学报*, 2015, 19(8): 81-87.
 - [7] YANG Y, DANG J, WANG Y. Medical image fusion method based on lifting wavelet transform and dual-channel PCNN [C]// 2014 IEEE 9th Conference on Industrial Electronics and Applications (ICIEA). IEEE, 2014: 1179-1182.
 - [8] AFONSO S B, EDGAR D, DUSTIN G M, et al. Certifying the Restricted Isometry Property is Hard[J]. *Information Theory IEEE Transactions on*, 2013, 59(6): 3448-3450.
 - [9] CANDÈS E J, TAO T. Decoding by Linear Programming[J]. *IEEE Transactions on Information Theory*, 2004, 54(4): 435-443.
 - [10] CANDÈS E J, ROMBERG J K, TAO T. Stable signal recovery from incomplete and inaccurate measurements[J]. *Communications on Pure & Applied Mathematics*, 2006, 59(8): 1207-1223.
 - [11] SAHOO S K, MAKUR A. Signal Recovery from Random Measurements via Extended Orthogonal Matching Pursuit[J]. *IEEE Transactions on Signal Processing*, 2015, 63(10): 2572-2581.
 - [12] FU Ying-hua. Reconstruction of compressed sensing and semi-QR factorization[J]. *Journal of Computer Applications*, 2008, 28(9): 2300-2302. (in Chinese)
傅迎华. 可压缩传感重构算法与近似 QR 分解[J]. *计算机应用*, 2008, 28(9): 2300-2302.
 - [13] LUO X, ZHANG J, DAI Q. A regional image fusion based on similarity characteristics[J]. *Signal Processing*, 2012, 92(5): 1268-1280.
 - [14] SHEN Xiao-hua, YANG Guo-sheng, ZHANG Huan-long. Improved on Approach of Image Fusion Based on Region-energy [J]. *Journal of Projectiles, Rockets, Missiles and Guidance*, 2006, 26(4): 279-281. (in Chinese)
申晓华, 杨国胜, 张焕龙. 改进的基于区域能量的图像融合方法[J]. *弹箭与制导学报*, 2006, 26(4): 279-281.
 - [15] XYDEAS C S, PETROVIC V. Objective image fusion performance measure[J]. *Electronics Letters*, 2000, 36(4): 308-309.
-
- (上接第 288 页)
- [2] HONKO P. Association Discovery from Relational Data via Granular Computing[J]. *Information Sciences*, 2013, 234(11): 136-149.
 - [3] TAGARELLI A. Exploring Dictionary-based Semantic Relatedness in Labeled Tree Data[J]. *Information Sciences An International Journal*, 2013, 220(1): 244-268.
 - [4] COZMAN F G. Independence for Full Conditional Probabilities: Structure, Factorization, Non-uniqueness, and Bayesian Networks[J]. *International Journal of Approximate Reasoning*, 2013, 54(9): 1261-1278.
 - [5] SHE Yan-hong. On the Rough Consistency Measures of Logic Theories and Approximate Reasoning in Rough Logic[J]. *International Journal of Approximate Reasoning*, 2014, 55(1): 486-499.
 - [6] YAN Lin, YAN Shuo. Granular Reasoning and Decision System's Decomposition[J]. *Journal of Software*, 2012, 7(3): 683-690.
 - [7] YAN Lin, LIU Qing. Researches on Granular Reasoning Based on Granular Space[C]// Proceedings of 2008 IEEE International Conference on Granular Computing. Hangzhou, China: 706-711.
 - [8] JIA Xiu-yi, LIAO Wen-he, TANG Zhen-min. Minimum Cost Attribute Reduction in Decision-theoretic Rough Set Models[J]. *Information Sciences*, 2013, 219: 151-167.
 - [9] MCALLISTER R A, ANGRYK R A. Abstracting for Dimensionality Reduction in Text Classification[J]. *International Journal of Intelligent Systems*, 2013, 28(2): 115-138.
 - [10] YAN Lin, SONG Jin-peng. Granular Trees Based on Different Data Sets and Their Modeling Applications[J]. *Computer Science*, 2014, 41(3): 258-262. (in Chinese)
闫林, 宋金鹏. 数据集的粒化树及其建模应用[J]. *计算机科学*, 2014, 41(3): 258-262.
 - [11] YAN Shuo, YAN Lin. Descriptive Method on Data Association Based on Granulation Trees[J]. *Pattern Recognition and Artificial Intelligence*, 2015, 28(12): 1057-1066. (in Chinese)
闫硕, 闫林. 数据关联的粒化树描述方法[J]. *模式识别与人工智能*, 2015, 28(12): 1057-1066.
 - [12] YAN Lin. Fundamentals of Mathematical Logic and Granular Computing[M]. Beijing: Science Press, 2007. (in Chinese)
闫林. 数理逻辑基础与粒计算[M]. 北京: 科学出版社, 2007.
 - [13] PEDRYCZ W. Granular Computing: Analysis and Design of Intelligent Systems[M]. Boca Raton: CRC Press/ Francis Taylor, 2013.
 - [14] KOLMAN B, BUSBY R C, ROSS S C. Discrete Mathematical Structures (fourth edition) [M]. New Jersey: Prentice-Hall, 2001.