数据合并的结构粒化方法与矩阵计算

闫 林[1] 高 伟[1] 闫 硕[2]
(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 为了研究数据合并问题,并使合并数据保持合并前的数据之间的关联关系,对各类数据信息给予了结构化的表示,对应产生了由数据集和加权关系组合构成的加权关联结构;进而通过数据集的合并粒化集,完成了加权关联结构向加权粒化结构的转换,使数据集中的数据依据粒化信息得到了合并,并保持或汇集了合并前的数据之间的关联信息,由此形成了数据合并的结构粒化方法。在此基础上,构建了加权关联矩阵和加权粒化矩阵,分别作为加权关联结构和加权粒化结构的矩阵表示。经中间变换和目标变换的矩阵计算,实现了加权关联矩阵向加权粒化矩阵的变换,产生了与结构粒化等价的矩阵变换方法,形成了程序设计的算法基础。

关键词 数据合并,合并粒化集,加权关联结构,加权粒化结构,加权关联矩阵,加权粒化矩阵

中图法分类号 TP181 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2017.09.049

Data Consolidation Based on Structure Granulation and Matrix Computation

YAN Lin[1] GAO Wei[1] YAN Shuo[2]
(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 study the problem of data consolidation, and make the consolidation data keep the associated information before consolidation, a structure composed of a data set together with a weighted relation was constructed, which refers to a weighted association structure. It demonstrates a structured representation of various data information. Then by using a granulation set, the weighted association structure is transformed into the weighted granulation structure. This makes the data in a granule to be consolidated according to the granulation information. At the same time, the consolidation data maintain or gather the associated information before consolidation, which leads to a structural granulation method of data consolidation. On this basis, two matrices were introduced, called the weighted association matrix and weighted granulation matrix which are matrix representations of the weighted association structure and weighted granulation structure respectively. Moreover, by matrix computations of elementary transformation and target transformation, the weighted association matrix is transformed into the weighted granulation matrix. This forms an equiva-lent way of data consolidation, and can be taken as the basis of programming algorithm.

Keywords Data consolidation, Granulation set, Weighted association structure, Weighted granulation structure, Weighted association matrix, Weighted granulation matrix

1 引言

信息科学的发展与数据研究的各个分支密切相关,同时各分支的成果又进一步推进了研究的深入,并伴随着新课题的出现。很多情况下,虽然不同分支关注的方向有所差异,但不同的研究往往交汇于共同的方面。比如,数据挖掘[1-4]、数据仓储[5-8]、数据推理[9-12]、数据约简[13-15]等数据研究的分支虽在方向上各有侧重,但数据的联合重组或合并归一问题常成为其研究中的共同关注点。同时由于数据研究的分支均与实际数据处理联系紧密,因此稍作留意就可看到实际中的数据重组或合并归一现象,可将其概括为相应模式,即一类数据联合重组成新的数据,或若干数据合并成同类数据的数据变化模式。为了认识该模式下的数据变化情况,不妨给予直观上的讨论分析。

在问题建模及程序设计过程中,研究者常把村庄、乡镇、成员、家庭、高校、院系、班级、学生等作为数据处理的对象,使它们融入模型构建和算法编程之中,以追求数据变化的程序化处理,以及事务管理的自动化系统。如果观察这些数据,则可看到联合重组的事实。比如:村庄组合成乡镇、成员组合成家庭、院系组合成高校、学生组合成班级等均是数据转换重组

到稿日期:2016-08-22 返修日期:2016-12-03 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,E-mail:hnsdyl@163.com;高 伟(1990-),男,硕士生,主要研究方向为粗糙集、数据挖掘,E-mail:gaoweihsd@foxmail.com;闫 硕(1987-),男,博士,主要研究方向为计算机代数、数据推理。

的例子。如何描述这种模式的转换是有意义的课题，也是问题程序化的前提。

此外，实际中还存在着同类数据合并归一的现象。比如若干企业合并成新的企业、不同植物嫁接成新的植物、两个天体吸引成大的天体、多条河流汇合成大的河流等均是合并归一的数据变化。与数据联合重组成新数据的转换重组不同，合并归一后，数据具有与之前相同的含义。此类实例还可枚举，比如：几所高校合并成高校、若干院系归整成院系、几处沙地沙化成沙漠、几块林地扩展成林区等。如何描述合并归一的数据变化也是值得考虑的数据处理问题。

无论是数据的联合重组，还是数据的合并归一，它们都展示了若干数据归为另一数据的变化，以下统称为数据合并处理，简称数据合并。由于数据合并与实际问题密切相关，因此对数据合并的研究具有理论和应用方面的意义。

数据合并意味着若干数据的整合或融合成的整体，这与粒计算[16-18]视一类数据为粒的数据聚类方式相一致。因此针对数据合并的讨论将借鉴粒计算的数据处理思路，以下工作可能提供粒计算研究的某种途径或处理手段。

数据合并是数据的重组或归一，对合并前数据关联信息的继承或汇集也是数据合并必须考虑的问题。仅关注数据自身的合并而忽视对关联信息的处理，将降低研究的意义，因为数据关联的信息必然传递给合并后的数据。

实际上，文献[19]已通过结构转换和矩阵变换的数学方法对数据合并问题进行了研究，同时也预留了继续探究的空间。比如合并数据的表示形式、关联信息的方向描述、数据关联的强弱程度、相关结论的严格证明等都是有待探究的问题，下面围绕这些问题展开研究。

2 合并数据的粒化表示

上文把数据的联合重组或合并归一称为数据合并，不妨把联合重组或合并归一后的新数据称为合并数据。对合并数据的表示是不可回避的问题，为此考虑数据集的概念。

设 U 是一集合，其元素称为数据，称 U 为数据集，此时 U 表示了一类数据的全体。比如 U 是高校学生的数据集、U 是若干企业的数据集、U 是一些村庄的数据集、U 是学校班级的数据集等，这些都反映了某类数据的聚集。

设 $x_1,x_2,\cdots,x_t(t\geqslant 1)$ 是数据集 U 中的一组数据，即 $x_i\in U(i=1,2,\cdots,t)$。合并 $x_1,x_2,\cdots,x_t$ 后，合并数据的表示形式可与子集 $\{x_1,x_2,\cdots,x_t\}$ 联系在一起，设想把 $\{x_1,x_2,\cdots,x_t\}$ 作为 $x_1,x_2,\cdots,x_t$ 联合重组或合并归一后的表示形式。$\{x_1\}$ 是 $t=1$ 的情况，这意味着 x_1 不与其他数据进行合并，将 $\{x_1\}$ 与 x_1 视为相同。因此如果把 U 中数据 $x_1,x_2,\cdots,x_t$ 的合并数据定义为 $\{x_1,x_2,\cdots,x_t\}$，U 中不参与合并的数据 x 用单数据集 $\{x\}$ 表示，则对合并数据的表示可与数据集 U 的分类相联系。

定义 1 设 U 是数据集，$E=\{E_1,E_2,\cdots,E_k\}$ 是 U 的子集构成的集合，即对于 $E_i\in E$，有 $E_i\subseteq U(i=1,2,\cdots,k)$。如果满足以下 3 条：

(1)$E_i\neq\varnothing(\mathrm{i}=1,2,\cdots,k)$；

(2)$E_i\cap E_j=\varnothing(i\neq j$ 且 $1\leqslant i,j\leqslant k)$；

(3)$E_1\cup E_2\cup\cdots\cup E_k=U$。

则称 $E=\{E_1,E_2,\cdots,E_k\}$ 是数据集 U 的合并粒化集，其中的元素 $E_i(i=1,2,\cdots,k)$ 称为粒。令 $E_i=\{x_1,x_2,\cdots,x_t\}$，当 $t>1$ 时，称粒 E_i 为数据 $x_1,x_2,\cdots,x_t$ 的合并数据或粒化表示；当 $t=1$ 时，$E_i=\{x_1\}$，此时称粒 E_i 为 x_1 的保留数据。

实际上，合并粒化集 $E=\{E_1,E_2,\cdots,E_k\}$ 是数据集 U 的划分[16]，按照粒计算[16-18]的理念，其也是对 U 的粒化，这种粒化可看做 U 中数据合并的一种方案。$E_i\subseteq U(i=1,2,\cdots,k)$ 表明，E_i 是数据集 U 中数据的子类，子类在粒计算研究中称为粒，因此这里对粒的定义一致于对粒的认识，与粒的数据归类特性相一致。同时，粒既指明了组合归一的数据，也用作合并数据的表示形式，形成了数据合并的粒化表示方法。

因此，如果数据集 U 的合并粒化集 $E=\{E_1,E_2,\cdots,E_k\}$ 有目的给定，那么粒 $E_1,E_2,\cdots,E_k$ 将提供数据集 U 中数据合并或保留的一种方案，并以粒 $E_1,E_2,\cdots,E_k$ 的形式给予合并数据的表示，展示了粒计算的数据处理思想。

3 加权关系和加权关联结构

数据集是相关数据的聚集，同时，数据之间并非孤立的，往往存在关联信息。如果整体考虑数据的聚类和数据关联信息，则可形成各类信息的结构化表示。因此，采用数学结构表示数据集和数据间的关联信息，通过结构转换完成数据合并，使合并数据继承或汇集合并前的关联信息是下面研究的课题，所以需要从数学结构的构造开始。

给定数据集 U，其对应产生笛卡尔积 $U\times U=\{\langle x,y\rangle\mid x,y\in U\}$[16]，其中的元素 $\langle x,y\rangle$ 称为序偶，记录了 x 和 y 之间的关联信息，明确了从 x 到 y 的方向或顺序，x 称为始数据，y 称为终数据。序偶 $\langle x,y\rangle$ 不同于序偶 $\langle y,x\rangle$，它们记录不同的关联顺序。笛卡尔积 $U\times U$ 由所有始数据和终数据均取自 U 的序偶构成。当 $S\subseteq U\times U$ 时，称 $U\times U$ 的子集 S 为 U 上的关系[16]，且可将 S 与权的概念相联系。

定义 2 设 S 是 U 上的关系，当满足 $x\in U$ 时，有 $\langle x,x\rangle\notin S$。引入如下概念：

(1)若 S 中每一序偶 $\langle x,y\rangle(\in S)$ 均被赋予一正数 w，则称 S 为 U 上的加权关系。把 $\langle x,y\rangle$ 和 w 组合记作 $\langle x,w,y\rangle$，仍采用 $\langle x,w,y\rangle\in S$ 的记法，$\langle x,w,y\rangle$ 称为加权边，x 称为始数据，y 称为终数据，w 称为 $\langle x,w,y\rangle$ 的权。

(2)将 U 和 U 上的加权关系 S 构成的整体记作 $W=(U,S)$，并称其为加权关联结构。

因此 S 既表示 U 上的关系，也表示 U 上的加权关系，可根据上下文判定区分。设 S 为 U 上的加权关系，当 $\langle x,w,y\rangle\in S$ 时，加权边 $\langle x,w,y\rangle$ 不仅记录了从 x 到 y 的有向关联，且通过权 w 反映了关联的紧密程度。定义 2 要求 $\langle x,x\rangle\notin S$，这是因为讨论不涉及数据自身的关联问题。

加权关联结构 $W=(U,S)$ 是两类数据信息 U 和 S 的结构化表示，不仅记录了数据集 U 中的数据信息，而且通过 S 中的加权边可了解数据的关联情况，并且可判定关联的紧密程度。“加权关联结构”中的“加权”表明加权关系 S 中每条边存在权值，这在文献[19]中并未涉及。将加权关系 S 融入 $W=(U,S)$ 中是因为仅考虑 U 中的数据不能反映各类信息的

全体。如公交系统:设 U 是公交站点的数据集,S 是 U 上的加权关系,对于 $\langle x,w,y\rangle\in S$,加权边 $\langle x,w,y\rangle$ 表示站点 x 和 y 相邻,且具有 w 的车流量。此时的加权边 $\langle x,w,y\rangle$ 反映了站点的关联及关联的程度。$W=(U,S)$ 记录了整个公交系统的信息。若仅有 U,不存在 S,即仅有站点,无车辆运行,那么公交系统将失去实际意义。

4 基于结构粒化的数据合并

设 $W=(U,S)$ 是加权关联结构,它是一类数据和数据之间有向关联及紧密程度的结构化表示。如果给定数据集 U 的合并粒化集 $E=\{E_1,E_2,\cdots,E_k\}$(见定义 1),则 $E_1,E_2,\cdots,E_k$ 是 U 中数据合并或保留的一种方案。

对于粒 $E_i,E_j\in E$ 且 $E_i\neq E_j$,若存在加权边 $\langle x,w,y\rangle\in S$,使得 $x\in E_i$ 且 $y\in E_j$,则称 $\langle E_i,E_j\rangle$ 为粒关联对,并称 $\langle x,w,y\rangle$ 为 $\langle E_i,E_j\rangle$ 的支撑,同时可能存在不同于 $\langle x,w,y\rangle$ 的加权边 $\langle x',w',y'\rangle\in S$,使得 $\langle x',w',y'\rangle$ 也是 $\langle E_i,E_j\rangle$ 的支撑,此时 $x'\in E_i$ 且 $y'\in E_j$。因此当 $\langle E_i,E_j\rangle$ 构成粒关联对时,除要求 $E_i\neq E_j$ 外,$\langle E_i,E_j\rangle$ 至少要有一个支撑。若存在 $t(t\geqslant 1)$ 条加权边:$\langle x_1,w_1,y_1\rangle\in S,\langle x_2,w_2,y_2\rangle\in S,\cdots,\langle x_t,w_t,y_t\rangle\in S$,使得加权边的序列 $\langle x_1,w_1,y_1\rangle,\langle x_2,w_2,y_2\rangle,\cdots,\langle x_t,w_t,y_t\rangle$ 包含了粒关联对 $\langle E_i,E_j\rangle$ 的所有支撑,则称该序列 $\langle x_1,w_1,y_1\rangle,\langle x_2,w_2,y_2\rangle,\cdots,\langle x_t,w_t,y_t\rangle$ 是粒关联对 $\langle E_i,E_j\rangle$ 的支撑组。令 $w^*=w_1+w_2+\cdots+w_t$,称 w^* 为粒关联对 $\langle E_i,E_j\rangle$ 的权,即 $\langle E_i,E_j\rangle$ 的权是其支撑组中各加权边的权相加的结果。

定义 3 设 $W=(U,S)$ 是加权关联结构,且 $E=\{E_1,E_2,\cdots,E_k\}$ 是数据集 U 的合并粒化集,令 $S^*=\{\langle E_i,w^*,E_j\rangle\mid\langle E_i,E_j\rangle$ 是粒关联对,且 w^* 是 $\langle E_i,E_j\rangle$ 的权$\}$,称 S^* 为 S 的加权粒化关系,其中 $\langle E_i,w^*,E_j\rangle$ 称为加权粒化边,E_i 称为始数据,E_j 称为终数据,w^* 称为 $\langle E_i,w^*,E_j\rangle$ 的权。

对于加权粒化关系 S^*,去除每一条加权粒化边 $\langle E_i,w^*,E_j\rangle$ 的权 w^* 后得到的关系仍记作 S^*,即 $S^*=\{\langle E_i,E_j\rangle\mid\langle E_i,w^*,E_j\rangle\in S^*\}$。此时 S^* 是 $E=\{E_1,E_2,\cdots,E_k\}$ 上的关系,即 $S^*\subseteq E\times E$。这可根据 S^* 是加权粒化关系或 S^* 是 E 上关系的表述进行区分。加权粒化关系 S^* 是对 E 上关系 S^* 中每一个粒关联对 $\langle E_i,E_j\rangle$ 指定权 w^* 后的结果。由定义 2(1)可知,加权粒化关系 S^* 是 E 上的加权关系,此时若令 $W^*=(E,S^*)$,由定义 2(2)可知,$W^*=(E,S^*)$ 是一个加权关联结构。

因此对于给定的加权关联结构 $W=(U,S)$,通过数据集 U 的合并粒化集 $E=\{E_1,E_2,\cdots,E_k\}$,可确定另一加权关联结构 $W^*=(E,S^*)$,从而通过合并粒化集 $E=\{E_1,E_2,\cdots,E_k\}$ 实现了从结构 $W=(U,S)$ 到结构 $W^*=(E,S^*)$ 的转换。

定义 4 结构 $W^*=(E,S^*)$ 称为加权关联结构 $W=(U,S)$ 的加权粒化结构。把通过 $E=\{E_1,E_2,\cdots,E_k\}$ 实现的从 $W=(U,S)$ 到 $W^*=(E,S^*)$ 的转换称为基于结构粒化的数据合并。

基于结构粒化的数据合并是结构 $W=(U,S)$ 粒化为结构 $W^*=(E,S^*)$ 的转换,是从加权关联结构开始,经给定的合并粒化集,使加权关系粒化为加权粒化关系后产生加权粒化结构的过程。这不仅完成了数据的合并,同时保持或汇集了有向关联的信息。为阐明这些事实,不妨进行如下分析。

(1)合并数据的分析:加权粒化结构 $W^*=(E,S^*)$ 中的数据集 E 是加权关联结构 $W=(U,S)$ 中数据集 U 的合并粒化集,合并粒化集 E 是 U 中哪些数据合并、哪些数据保留的方案?从 $W=(U,S)$ 到 $W^*=(E,S^*)$ 的转换完成了 U 中数据的合并或保留,E 中的粒明确了合并数据或保留数据的形式,并作为 $W^*=(E,S^*)$ 的数据集得到了记录。

(2)有向联系得以保持或汇集的分析:对于结构 $W=(U,S)$ 和 $W^*=(E,S^*)$,设 $\langle E_i,w^*,E_j\rangle\in S^*$,则存在 $\langle x_1,w_1,y_1\rangle\in S,\langle x_2,w_2,y_2\rangle\in S,\cdots,\langle x_t,w_t,y_t\rangle\in S$,使得 $\langle x_1,w_1,y_1\rangle,\langle x_2,w_2,y_2\rangle,\cdots,\langle x_t,w_t,y_t\rangle(t\geqslant 1)$ 是粒关联对 $\langle E_i,E_j\rangle$ 的支撑组,此时 $x_k\in E_i$ 且 $y_k\in E_j(k=1,2,\cdots,t)$。边 $\langle x_k,w_k,y_k\rangle(k=1,2,\cdots,t)$ 与边 $\langle E_i,w^*,E_j\rangle$ 通过 $x_k\in E_i$ 且 $y_k\in E_j$ 产生联系,加权粒化边 $\langle E_i,w^*,E_j\rangle$ 记录的从 E_i 到 E_j 的关联是对其支撑组记录的从 x_1 到 y_1,从 x_2 到 y_2,…,从 x_t 到 y_t 的所有关联的保持或汇集,因为边 $\langle x_k,w_k,y_k\rangle(k=1,2,\cdots,t)$ 记录的从 x_k 到 y_k 的关联通过 $x_k\in E_i$ 且 $y_k\in E_j$ 传递给了边 $\langle E_i,w^*,E_j\rangle$。当 $t=1$ 时,序列 $\langle x_1,w_1,y_1\rangle,\langle x_2,w_2,y_2\rangle,\cdots,\langle x_t,w_t,y_t\rangle$ 就是 $\langle x_1,w_1,y_1\rangle$,此时加权粒化边 $\langle E_i,w^*,E_j\rangle$ 记录的关联信息是对 $\langle x_1,w_1,y_1\rangle$ 记录关联的保持;当 $t>1$ 时,$\langle E_i,w^*,E_j\rangle$ 记录的数据关联是支撑组 $\langle x_1,w_1,y_1\rangle,\langle x_2,w_2,y_2\rangle,\cdots,\langle x_t,w_t,y_t\rangle$ 记录关联的汇集,并通过 $w^*=w_1+w_2+\cdots+w_t$ 予以体现。实际上,表达式 $w^*=w_1+w_2+\cdots+w_t$ 确定的权值运算方式源于具体问题,比如:企业 A 给企业 C 供货 30%,企业 B 给企业 C 供货 40%,则 A 和 B 合并后给 C 的供货量是 30%+40%=70%。

因此,基于结构粒化的数据合并通过结构的转换完成了数据的合并,且保持或汇集了数据关联的信息,形成了数据合并的结构粒化方法。尽管该方法着重理论上的分析,但为程序化方法的建立提供了基础,从而引出如下的讨论。

5 矩阵表示与矩阵变换

5.1 加权关联矩阵和加权粒化矩阵

定义 5 设 $W=(U,S)$ 是加权关联结构,其中 $U=\{d_1,d_2,\cdots,d_m\}$,$S=\{e_1,e_2,\cdots,e_n\}$ 且 $e_k=\langle d_{k1},w_k,d_{k2}\rangle(k=1,2,\cdots,n)$。构造 $m+1$ 行 n 列的矩阵 $M(W)$:

$$M(W)=(r_{ij})_{(m+1)\times n}$$

其中,$r_{ij}(i=1,\cdots,m+1;j=1,\cdots,n)$ 的取值规定为:当 $i=1$ 时,$r_{ij}=w_j(j=1,\cdots,n)$,即第一行是由 n 条加权边 $e_1,e_2,\cdots,e_n$ 的权构成的行向量 $(w_1,w_2,\cdots,w_n)$,称为权值行向量;当 $2\leqslant i\leqslant m+1$ 时,若 d_{i-1} 是 e_j 的始数据,则 $r_{ij}=1$;若 d_{i-1} 是 e_j 的终数据,则 $r_{ij}=-1$;若 d_{i-1} 既非 e_j 的始数据,也非 e_j 的终数据,则 $r_{ij}=0$。称 $M(W)$ 为 $W=(U,S)$ 的加权关联矩阵,其每一行称为行向量,每一列称为列向量。

例 1 设 $W=(U,S)$ 是加权关联结构,其中 $U=\{1,2,3,4,5\}$,$S=\{e_1,e_2,e_3,e_4,e_5,e_6\}$,$e_1=\langle 1,0.2,2\rangle$,$e_2=\langle 1,0.5,5\rangle$,$e_3=\langle 1,0.3,3\rangle$,$e_4=\langle 2,0.1,4\rangle$,$e_5=\langle 3,0.4,5\rangle$,$e_6=\langle 4,

0.1,5〉。则 $W=(U,S)$ 的加权关联矩阵 $M(W)$ 如式(1)所示：

$$\begin{array}{c} \\ \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array}\begin{array}{c} \begin{array}{cccccc} e_1 & e_2 & e_3 & e_4 & e_5 & e_6 \end{array} \\ \begin{bmatrix} 0.2 & 0.5 & 0.3 & 0.1 & 0.4 & 0.1 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 & 0 & 1 \\ 0 & -1 & 0 & 0 & -1 & -1 \end{bmatrix} \end{array} \tag{1}$$

其中，第一行(0.2,0.5,0.3,0.1,0.4,0.1)是权值行向量，数据 $i(i=1,2,3,4,5)$ 对应的行向量由 1，−1 或 0 构成，加权边 $e_k(k=1,2,3,4,5,6)$ 对应的列向量的第一分量是权值，其余分量中有一个 1 和一个 −1，其他都是 0。

加权关联结构 $W=(U,S)$ 可确定加权关联矩阵 $M(W)$。反之，加权关联矩阵 $M(W)$ 中的数值信息与行列标记完全可以确定数据集 U 和加权关系 S，从而产生加权关联结构 $W=(U,S)$。因此 $W=(U,S)$ 和 $M(W)$ 相互唯一确定。矩阵是可以程序化的数据形式，因此加权关联结构 $W=(U,S)$ 的矩阵表示提供了该结构的程序化处理方法。

定义 6 设 $W^*=(E,S^*)$ 是 $W=(U,S)$ 的加权粒化结构，$E=\{E_1,E_2,\cdots,E_k\}$ 且 $S^*=\{e_1{}^*,e_2{}^*,\cdots,e_q{}^*\}$，其中 $e_l^*=\langle E_{l_1},w_l^*,E_{l_2}\rangle(l=1,2,\cdots,q)$。构造 $k+1$ 行 q 列矩阵 $M(W)^*$：

$M(W)^*=(r_{ij}^*)_{(k+1)\times q}$

其中，$r_{ij}^*(i=1,\cdots,k+1;j=1,\cdots,q)$ 的取值规定为：当 $i=1$ 时，$r_{ij}^*=w_j^*(j=1,\cdots,q)$，即第一行是 q 条加权粒化边 e_1^*，e_2^*，$\cdots$，e_q^* 的权值构成的行向量$(w_1^*,w_2^*,\cdots,w_q^*)$，称为权值行向量；当 $2\leqslant i\leqslant k+1$ 时，若 E_{i-1} 是 e_j^* 的始数据，则 $r_{ij}^*=1$；若 E_{i-1} 是 e_j^* 的终数据，则 $r_{ij}^*=-1$；若 E_{i-1} 既非 e_j^* 的始数据，也非 e_j^* 的终数据，则 $r_{ij}^*=0$。称 $M(W)^*$ 为 $W^*=(E,S^*)$ 的加权粒化矩阵，其每一行称为行向量，每一列称为列向量。

显然，加权粒化结构 $W^*=(E,S^*)$ 与加权粒化矩阵 $M(W)^*$ 之间也相互唯一确定。

例 2 考虑例 1 中的加权关联结构 $W=(U,S)$，其中 $U=\{1,2,3,4,5\}$，$S=\{e_1,e_2,e_3,e_4,e_5,e_6\}$，$e_1=\langle 1,0.2,2\rangle$，$e_2=\langle 1,0.5,5\rangle$，$e_3=\langle 1,0.3,3\rangle$，$e_4=\langle 2,0.1,4\rangle$，$e_5=\langle 3,0.4,5\rangle$，$e_6=\langle 4,0.1,5\rangle$。给定 U 的合并粒化集 $E=\{\{1,2\},\{3,4\},\{5\}\}$，则通过 E 可得到 $W=(U,S)$ 的加权粒化结构 $W^*=(E,S^*)$，其中的数据集为 $E=\{\{1,2\},\{3,4\},\{5\}\}$，由定义 3 可得到加权粒化关系 $S^*=\{e_1^*,e_2^*,e_3^*\}$ 且 $e_1^*=\langle\{1,2\},0.5,\{5\}\rangle$，$e_2^*=\langle\{1,2\},0.4,\{3,4\}\rangle$，$e_3^*=\langle\{3,4\},0.5,\{5\}\rangle$(将分析产生过程)。$W^*$ 的加权粒化矩阵 $M(W)^*$ 如式(2)所示：

$$\begin{array}{c} \\ \\ \{1,2\} \\ \{3,4\} \\ \{5\} \end{array}\begin{array}{c} \begin{array}{ccc} e_1^* & e_2^* & e_3^* \end{array} \\ \begin{bmatrix} 0.5 & 0.4 & 0.5 \\ 1 & 1 & 0 \\ 0 & -1 & 1 \\ -1 & 0 & -1 \end{bmatrix} \end{array} \tag{2}$$

现对加权粒化边 e_1^*，e_2^* 和 e_3^* 的形成进行分析：

(1)$e_1^*=\langle\{1,2\},0.5,\{5\}\rangle$ 的产生缘于 S 的加权边 e_1,e_2,e_3,e_4,e_5,e_6，只有 $e_2=\langle 1,0.5,5\rangle$ 是粒关联对 $\langle\{1,2\},\{5\}\rangle$ 的支撑，满足 $1\in\{1,2\}$ 且 $5\in\{5\}$。此时 $\langle\{1,2\},\{5\}\rangle$ 的支撑组是 $e_2=\langle 1,0.5,5\rangle$，且 $\langle\{1,2\},\{5\}\rangle$ 的权等于 0.5，由此形成加权粒化边 $e_1^*=\langle\{1,2\},0.5,\{5\}\rangle$。

(2)$e_2^*=\langle\{1,2\},0.4,\{3,4\}\rangle$ 的形成缘于 $e_3=\langle 1,0.3,3\rangle$，$e_4=\langle 2,0.1,4\rangle$ 是粒关联对 $\langle\{1,2\},\{3,4\}\rangle$ 的支撑组，满足 $1\in\{1,2\}$ 且 $3\in\{3,4\}$，以及 $2\in\{1,2\}$ 且 $4\in\{3,4\}$，此时 $\langle\{1,2\},\{3,4\}\rangle$ 的权等于 $0.3+0.1=0.4$，因此形成加权粒化边 $e_2^*=\langle\{1,2\},0.4,\{3,4\}\rangle$。

(3)$e_3^*=\langle\{3,4\},0.5,\{5\}\rangle$ 的形成缘于 $e_5=\langle 3,0.4,5\rangle$，$e_6=\langle 4,0.1,5\rangle$ 是粒关联对 $\langle\{3,4\},\{5\}\rangle$ 的支撑组，满足 $3\in\{3,4\}$ 且 $5\in\{5\}$，以及 $4\in\{3,4\}$ 且 $5\in\{5\}$，此时 $\langle\{3,4\},\{5\}\rangle$ 的权等于 $0.4+0.1=0.5$，由此产生加权粒化边 $e_3^*=\langle\{3,4\},0.5,\{5\}\rangle$。

注释说明：对于例 1 中的加权边 $e_1=\langle 1,0.2,2\rangle$ 和例 2 中的合并粒化集 $E=\{\{1,2\},\{3,4\},\{5\}\}$，虽然 $1\in\{1,2\}$ 且 $2\in\{1,2\}$，但 $\langle\{1,2\},\{1,2\}\rangle$ 不是粒关联对。因为对粒关联对 $\langle E_i,E_j\rangle$ 有 $E_i\neq E_j$ 的要求，所以加权边 $e_1=\langle 1,0.2,2\rangle$ 与任何加权粒化边均无联系，它是冗余的。

5.2 中间变换和目标变换

设 $W=(U,S)$ 是加权关联结构，其加权关联矩阵 $M(W)$ 的第一行是权值行向量$(w_1,w_2,\cdots,w_n)$，其他行向量由 1，−1或 0 构成。这些行向量的运算与数据合并密切相关，为此需要考虑行向量的相加运算。行向量的相加与权值行向量$(w_1,w_2,\cdots,w_n)$无关，是除权值行向量的行向量相加。

考虑 $M(W)$ 中的有限个行向量$(r_{i_1},r_{i_2},\cdots,r_{i_n})$，$(r_{j_1},r_{j_2},\cdots,r_{j_n})$，$\cdots$，$(r_{k_1},r_{k_2},\cdots,r_{k_n})$，定义它们的相加为：$(r_{i_1},r_{i_2},\cdots,r_{i_n})+(r_{j_1},r_{j_2},\cdots,r_{j_n})+\cdots+(r_{k_1},r_{k_2},\cdots,r_{k_n})=(r_{i_1}+r_{j_1}+\cdots+r_{k_1},r_{i_2}+r_{j_2}+\cdots+r_{k_2},\cdots,r_{i_n}+r_{j_n}+\cdots+r_{k_n})$，即有限个行向量的相加由对应分量相加而确定，其结果仍是由 1，−1 或 0 构成的行向量。

设 $M(W)$ 是 $W=(U,S)$ 的加权关联矩阵，并设 $U=\{d_1,d_2,\cdots,d_m\}$ 且 $S=\{e_1,e_2,\cdots,e_n\}$。给定 U 的合并粒化集 $E=\{E_1,E_2,\cdots,E_k\}$，对于 $E_i\in E(i=1,2,\cdots,k)$，令 $E_i=\{d_{i_1},d_{i_2},\cdots,d_{i_t}\}(\subseteq U;t\geqslant 1)$，针对 $M(W)$ 可进行如下操作：(1)在 $M(W)$ 中将粒 $E_i=\{d_{i_1},d_{i_2},\cdots,d_{i_t}\}$ 中数据 $d_{i_1},d_{i_2},\cdots,d_{i_t}$ 对应的行向量相加所得到的行向量称为 E_i-行向量$(i=1,2,\cdots,k)$。

(1)在 $M(W)$ 中将粒 $E_i=\{d_{i_1},d_{i_2},\cdots,d_{i_t}\}$ 中数据 $d_{i_1},d_{i_2},\cdots,d_{i_t}$ 对应的行向量相加所得到的行向量称为 E_i-行向量$(i=1,2,\cdots,k)$。

(2)将矩阵 $M(W)$ 第一行的权值行向量作为第 1 行，E_1-行向量作为第 2 行，E_2-行向量作为第 3 行，$\cdots$，E_k-行向量作为第 $k+1$ 行的矩阵记作 $M_E(W)$，并用 E_i 对 E_i-行向量进行标记，仍用 $e_1,e_2,\cdots,e_n$ 对 $M_E(W)$ 的列向量进行标记，$M_E(W)$ 称为加权合并矩阵。

中间变换：从加权关联矩阵 $M(W)$ 到加权合并矩阵 $M_E(W)$ 的变换称为中间变换。

对于数据集 U 的合并粒化集 $E=\{E_1,E_2,\cdots,E_k\}$，及 $E_i\in E(i=1,2,\cdots,k)$，E_i-行向量就是 E_i 中数据对应的 $M(W)$ 中行向量相加的结果。中间变换是通过合并粒化集 E 完成的从 $M(W)$ 到 $M_E(W)$ 的变换。当粒 $E_i=\{d\}$ 时，数据 d 在

$M(W)$中对应的行向量与 $M_E(W)$中的 E_i-行向量相同,且 $M(W)$和 $M_E(W)$的第一行相同,都是 $M(W)$的权值行向量。

例 3　对于式(1)中的加权关联矩阵 $M(W)$,通过合并粒化集 $E=\{\{1,2\},\{3,4\},\{5\}\}$,可以得到加权合并矩阵 $M_E(W)$。具体地,$M_E(W)$的第一行就是 $M(W)$第一行的权值行向量(0.2,0.5,0.3,0.1,0.4,0.1)。分析 $M_E(W)$中的其他行向量:在式(1)的 $M(W)$中,数据 1 对应的行向量为(1,1,1,0,0,0),数据 2 对应的行向量为(−1,0,0,1,0,0),数据 3 对应的行向量为(0,0,−1,0,1,0),数据 4 对应的行向量为(0,0,−1,0,1),数据 5 对应的行向量为(0,−1,0,0,−1,−1)。令 $E_1=\{1,2\}$,$E_2=\{3,4\}$,$E_3=\{5\}$,则 E_1-行向量是数据 1 和 2 对应行向量相加的结果:(1,1,1,0,0,0)+(−1,0,0,1,0,0)=(0,1,1,1,0,0);E_2-行向量是数据 3 和 4 对应行向量相加的结果:(0,0,−1,0,1,0)+(0,0,0,−1,0,1)=(0,0,−1,−1,1,1);E_3-行向量与数据 5 对应的行向量(0,−1,0,0,−1,−1)相同,因为粒 $E_3=\{5\}$由单数据 5 构成,所以由权值行向量(0.2,0.5,0.3,0.1,0.4,0.1),E_1-行向量(0,1,1,1,0,0),E_2-行向量(0,0,−1,−1,1,1)和 E_3-行向量(0,−1,0,0,−1,−1)构成加权合并矩阵 $M_E(W)$,如式(3)所示:

$$\begin{array}{c} \\ \\ \{1,2\} \\ \{3,4\} \\ \{5\} \end{array}\begin{array}{c} \begin{array}{cccccc} e_1 & e_2 & e_3 & e_4 & e_5 & e_6 \end{array} \\ \begin{bmatrix} 0.2 & 0.5 & 0.3 & 0.1 & 0.4 & 0.1 \\ 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & -1 & -1 & 1 & 1 \\ 0 & -1 & 0 & 0 & -1 & -1 \end{bmatrix} \end{array} \tag{3}$$

其中,$\{1,2\}(=E_1)$,$\{3,4\}(=E_2)$和$\{5\}(=E_3)$分别用于对 E_1-行向量,E_2-行向量和 E_3-行向量的标记,e_1,e_2,e_3,e_4,e_5,e_6 仍然分别用于对各列向量的标记。

因此,利用合并粒化集 $E=\{\{1,2\},\{3,4\},\{5\}\}$中各粒包含的数据信息,可以实现加权关联矩阵 $M(W)$到加权合并矩阵 $M_E(W)$的中间变换。中间变换意味着 $M_E(W)$并非加权粒化矩阵 $M(W)^*$,仍需要对 $M_E(W)$实施进一步的变换。

定义 7　设 $M_E(W)$是加权合并矩阵,对于 e_i 对应的列向量,如果除权值分量外其他分量全为 0,则称 e_i 对应的列向量为列零向量;对于 $e_{j_1},e_{j_2},\cdots,e_{j_r}(r\geqslant 1)$对应的列向量,它们都不是列零向量,如果除权值分量外,这些列向量的其他对应分量都相等,且无另外满足该条件的列向量,则称 $e_{j_1},e_{j_2},\cdots,e_{j_r}$ 对应的列向量为平行列组。

针对加权合并矩阵 $M_E(W)$,可实施如下操作:

(1)删除 $M_E(W)$中的列零向量。

(2)合并 $M_E(W)$中的平行列组。该操作如此约定:当 $e_{j_1},e_{j_2},\cdots,e_{j_r}$ 对应的列向量是平行列组时,将该平行列组合并成一个列向量,用$\{e_{j_1},e_{j_2},\cdots,e_{j_r}\}$标记,其第一分量是权值 w^* 且 $w^*=w_{j_1}+w_{j_2}+\cdots+w_{j_r}$,其中 w_{ji} 是 e_{j_i} 的权($i=1,2,\cdots,r$);其他分量与该平行列组中每一列向量的对应分量相同。当 $r=1$ 时,$\{e_{j_1},e_{j_2},\cdots,e_{j_r}\}$变为$\{e_{j_1}\}$,$\{e_{j_1}\}$表示的列向量与 e_{j_1} 对应的列向量相同,即当一条加权边 e_{j_1} 构成平行列组时,平行列组的合并用$\{e_{j_1}\}$标记,并与 e_{j_1} 相同。

目标变换:删去 $M_E(W)$的所有列零向量,合并 $M_E(W)$中的每一平行列组,得到的矩阵记作 $M_E(W)'$,将其称为加权目标矩阵。从 $M_E(W)$到 $M_E(W)'$的变换称为目标变换。

例 4　对于式(3)中的加权合并矩阵 $M_E(W)$,通过目标变换,可求得 $M_E(W)'$:在式(3)中 e_1 对应的列向量$(0.2,0,0,0)^{\mathrm{T}}$ 是列零向量,按照目标变换的约定应将其删去;在式(3)中 e_3 和 e_4 对应的列向量$(0.3,1,-1,0)^{\mathrm{T}}$ 和$(0.1,1,-1,0)^{\mathrm{T}}$ 是平行列组,合并该平行列组得到列向量$(0.4,1,-1,0)^{\mathrm{T}}$,将其标记为$\{e_3,e_4\}$,其权值 0.4 等于 e_3 的权值 0.3 和 e_4 的权值 0.1 之和,其他分量与 e_3 或 e_4 对应的分量相同;在式(3)中 e_5 和 e_6 对应的列向量$(0.4,0,1,-1)^{\mathrm{T}}$ 和$(0.1,0,1,-1)^{\mathrm{T}}$ 是平行列组,该平行列组的合并产生列向量$(0.5,0,1,-1)^{\mathrm{T}}$,标记为$\{e_5,e_6\}$;在式(3)中 e_2 对应的列向量$(0.5,1,0,-1)^{\mathrm{T}}$ 自身是平行列组,合并后仍是$(0.5,1,0,-1)^{\mathrm{T}}$,标记为$\{e_2\}$。由此便得到加权目标矩阵 $M_E(W)'$,如式(4)所示:

$$\begin{array}{c} \\ \\ \{1,2\} \\ \{3,4\} \\ \{5\} \end{array}\begin{array}{c} \begin{array}{ccc} \{e_2\} & \{e_3,e_4\} & \{e_5,e_6\} \end{array} \\ \begin{bmatrix} 0.5 & 0.4 & 0.5 \\ 1 & 1 & 0 \\ 0 & -1 & 1 \\ -1 & 0 & -1 \end{bmatrix} \end{array} \tag{4}$$

其中,行向量的标记与式(3)相同,从而实现了从 $M_E(W)$到 $M_E(W)'$的目标变换。

比较式(2)中的加权粒化矩阵 $M(W)^*$ 和式(4)中的加权目标矩阵 $M_E(W)'$后可知它们完全相同。这并非偶然,在任何情况下,加权目标矩阵 $M_E(W)'$ 就是加权粒化矩阵 $M(W)^*$,因此可总结为如下的结论。

5.3　主要结论

设 $W=(U,S)$是一个加权关联结构,给定 U 的合并粒化集 $E=\{E_1,E_2,\cdots,E_k\}$,通过 E 可以实现从 $W=(U,S)$到加权粒化结构 $W^*=(E,S^*)$的转换。

同时由 $W=(U,S)$可确定其加权关联矩阵 $M(W)$,由 $W^*=(E,S^*)$可确定其加权粒化矩阵 $M(W)^*$。此时,从 $M(W)$出发,经中间变换得到加权合并矩阵 $M_E(W)$,再经目标变换产生加权目标矩阵 $M_E(W)'$。则如下结论成立:

$$M_E(W)'=M(W)^*$$

限于篇幅,略去该结论的证明,可将例 1—例 4 中的相关矩阵看作对该结论的验证。

值得说明的是:从 $W=(U,S)$经 $E=\{E_1,E_2,\cdots,E_k\}$到 $W^*=(E,S^*)$的转换是基于结构粒化的数据合并,由于 $W=(U,S)$和 $M(W)$,以及 $W^*=(E,S^*)$和 $M(W)^*$ 相互确定,因此用结构转换实现的数据合并可通过矩阵变换等价完成。矩阵变换是从 $M(W)$开始,经中间变换和目标变换,得到目标矩阵 $M_E(W)'(M_E(W)'=M(W)^*)$的代数运算。由于矩阵变换可编程完成,因此矩阵变换方法使数据合并的程序化成为可能。

结束语　基于结构粒化和矩阵变换的数据合并的讨论,建立了数据合并的两种等价方法,前者集中于理论分析,后者建立了变换运算,一个侧重理论,一个旨在编程。与文献[19]的讨论相比,本文明确给出了数据合并的定义,并将数据关联的方向、关联程度的强弱、关联信息的继承、关联合并的汇集等以前未涉及的问题进行了详尽的讨论,并给出了相应的处理方法。同时,给出了结论 $M_E(W)'=M(W)^*$ 的证明,但因篇幅有限,未列出较长证明的过程。因此上述的讨论更为深入,表述也更为细腻。

观察上述结构转换和矩阵变换的数据合并方法可知,合

(下转第 299 页)

nese Inertial Technology,2011,19(6):654-658.(in Chinese)
郝燕玲,杨峻巍,陈亮,等.基于 NPF-CKF 的捷联惯导系统动基座初始对准技术[J].中国惯性技术学报,2011,19(6):654-658.

[4] GE Q,LI W,WEN C. SCKF-STF-CN:a universal nonlinear filter for maneuver target tracking[J]. Journal of Zhejiang University SCIENCE C,2011,12(8):678-686.

[5] ZHANG Y,RUI G S,MIAO J,et al. Location Technology Based on the Extend Cubature Kalman Filter[J]. Opto-Electronic Engineering,2012,39(4):37-43.(in Chinese)
张洋,芮国胜,苗俊,等.扩展容积卡尔曼滤波定位技术研究[J].光电工程,2012.39(4):37-43.

[6] GE Q B,LI W B,SUN R Y,et al. Research on centralized fusion algorithms based on EKF for multisensor non-linear systems [J]. Acta Automatica Sinica,2013,39(6):816-825.

[7] CUI P Y,ZHENG L F,PEI F J,et al. Study on integrated navigation method based on self adjusting particle filter[J]. Computer Engineering,2008,34(14):185-187.(in Chinese)
崔平远,郑黎方,裴福俊,等.基于自调整粒子滤波的组合导航方法研究[J].计算机工程,2008,34(14):185-187.

[8] QIN Z. Improved particle filter and its application in GPS dynamic positioning[J]. Global Positioning System,2010,35(5):25-28.(in Chinese)
秦臻.改进的粒子滤波及其在 GPS 动态定位中的应用[J].全球定位系统,2010,35(5):25-28.

[9] FENG C,WANG M,JI Q B,et al. Analysis and comparison of particle filter resampling algorithm [J]. Journal of system simulation,2009,21(4):1101-1105.(in Chinese)
冯驰,王萌,汲清波,等.粒子滤波器重采样算法的分析与比较[J].系统仿真学报,2009,21(4):1101-1105.

[10] 朱志宇.粒子滤波算法及其应用[M].北京:科学出版社,2010.

[11] GORDON N J,SALMOND D J,SMITH A F M,et al. Novel approach to nonlinear/non-Gaussian Bayesian state estimation[J]. IEE Proceedings F-Radar and Signal Processing, IET, 1993, 140(2):107-113.

[12] SPALL J C. Estimation via markov chain monte carlo[J]. IEEE Control Systems,2003,23(2):34-45.

[13] ANDRIEU C,DJURIĆ P M,DOUCET A,et al. Model selection by MCMC computation [J]. Signal Processing, 2001, 81 (1): 19-37.

[14] GODSILL S,CLAPP T. Improvement strategies for Monte Carlo particle filters[M]// Sequential Monte Carlo Methods in Practice. Springer New York,2001:139-158.

(上接第 265 页)

并粒化集是加权粒化关系和加权粒化结构产生的基础。由于合并粒化集的构成与粒计算研究者的数据处理思想一致,因此上述针对数据合并的研究也可看作粒计算研究的一种途径。

参 考 文 献

[1] ZHANG J B,LI T R,CHEN H M. Composite rough sets for dynamic data mining[J]. Information Sciences,2014,257:81-100.

[2] ZHANG J B,LI T R,RUAN D. Neighborhood rough sets for dynamic data mining [J]. International Journal of Intelligent Systems,2012,27(4):317-342.

[3] HONKO P. Association discovery from relational data via granular computing [J]. Information Sciences, 2013, 234 (11): 136-149.

[4] MERIGO J M. The probabilistic weighted average and its application in multiperson decision making[J]. International Journal of Intelligent Systems,2012,27(5):457-476.

[5] BEAUBOUEF T,PETRY F. Fuzzy rough set techniques for uncertainty processing in a relational database[J]. International Journal of Intelligent System,2000,15(5):389-424.

[6] BEAUBOUEF T,PETRY F,ARORA G. Information-theoretic measures of uncertainty for rough sets and rough relational database[J]. Information Sciences,1998,109(1-4):185-195.

[7] 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.

[8] TAGARELLI A. Exploring dictionary-based semantic relatedness in labeled tree data[J]. Information Sciences,2013,220(1):244-268.

[9] SHE Y L. 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.

[10] YAN S,YAN L,WU J Z. Rough data-deduction based on the upper approximation [J]. Information Sciences,2016,373:308-320.

[11] YAN L,YAN S. Granular reasoning and decision systems decomposition [J]. Journal of Software,2012,7(3):683-690.

[12] YAN L,YAN S. Researches on rough truth of rough axioms based on granular reasoning [J]. Journal of Software, 2014, 9(2):265-273.

[13] LI J H,MEI C L,LV Y J. Incomplete decision contexts:approximate concept construction, rule acquisition and knowledge reduction [J]. International Journal of Approximate Reasoning, 2013,54(1):149-165.

[14] JIA X Y,LIAO W H,TANG Z M. Minimum cost attribute reduction in decision-theoretic rough set models [J]. Information Sciences,2013,219(1):151-167.

[15] MCALLISTER R A,ANGRYK R A. Abstracting for dimensionality reduction in text classification [J]. International Journal of Intelligent Systems,2013,28(2):115-138.

[16] 闫林.数理逻辑基础与粒计算[M].北京:科学出版社,2007.

[17] PEDRYCZ W. Granular computing:analysis and design of intelligent systems [M] . Boca Raton,USA:CRC Press Francis Taylor,2013.

[18] LI J H,MEI C L,XU W H,et al. Concept learning via granular computing: A cognitive viewpoint [J]. Information Sciences, 2015,298(1):447-467.

[19] YAN L,LIU T,YAN S,et al. Data combination method based on structure's granulation[J]. Journal of Computer Applications,2015,35(2):358-363.(in Chinese)
闫林,刘涛,闫硕,等.基于结构粒化的数据合并方法[J].计算机应用,2015,35(2):358-363.