基于粗糙集和形式概念分析的属性约简研究

马 坤

（大连海事大学 信息科学技术学院，辽宁省大连市 中国 116026）

摘 要 形式概念分析是一种数据分析和知识表达的有力工具，它以形式背景为基础。一个形式背景由对象，属性及对象与属性之间的二元关系构成。形式概念用集值算子描述，是对象子集和属性子集构成的二元组。作为知识处理的重要工具,形式背景的属性约简理论已被应用到很多领域。本文对属性约简方法的研究集中在利用粗糙集、粒计算和形式概念分析等基本方法，指明其发展趋势与数据的动态性、智能算法之间的相互融合密切相关。本文仅针对属性约简算法之间的融合发展进行简要论述，未对其进行更深入探讨，多种属性约简算法的融合研究是属性约简算法的发展趋势。

关键词 粗糙集 粒计算 形式概念分析

中图法分类号 TP311 **文献标识码** A

Research on Attribute Reduction Based on Rough Set and Formal Concept Analysis

Ma kun

(School of Information Science and Technology, Dalian maritime university, Liaoning Dalian 116026 China)

Abstract Formal concept analysis is a powerful tool for data analysis and knowledge expression, which is based on formal context. A formal context consists of objects, attributes and the binary relationship between objects and attributes. Formal concept is described by set-valued operator, which is a binary composed of object subset and attribute subset. As an important tool of knowledge processing, attribute reduction theory of formal context has been applied to many fields. In this paper, the research of attribute reduction methods focuses on the use of rough set, granular computing and formal concept analysis, and points out that its development trend is closely related to the dynamic of data and the fusion of intelligent algorithms. In this paper, the fusion development of attribute reduction algorithms is briefly discussed, and it is not further discussed. The fusion research of multiple attribute reduction algorithms is the development trend of attribute reduction algorithm.

Key words Rough set; Granular computing; Formal concept analysis

1 引言

属性约简被用于数据挖掘的预处理环节，是重要的特征选取方法之一[1]，指在保持系统分类不变的前提下，删除冗余属性[2]，得到有效规则库，从而辅助决策，其数学描述为：设 K = (U，R) 为一个知识库，在非空有限论域 U 上，R 为等价关系集合，K 中所定义的所有等价关系的族记作 IND(K)。若 P ⊆ R 是满足 IND(K) = IND (R) 的极小属性子集，则称 P 为 R 的属性约简。

大数据时代，众多应用领域的数据量急剧增加，数据分析前需降低属性维数，属性约简方法便成为一项重要的基础性工作。Wong 等已经证明寻找最小约简是 NP-Hard 问题[3]，说明属性约简仍然是具有挑战性的研究议题。因此，本研究结合属性约简方法相关文献的关键词词频分析结果，梳理属性约简的基本方法，总结属性约简方法的未来发展趋势。

2 关键词分析

“属性约简”、“Attribute Reduction”为关键词，分别在 CNKI 及 Web of Science 数据库中检索 2010 年-2019 年的文献共计 1341 篇，其中中文文献 1 113 篇，英文文献 228 篇，以上述文献的关键词作为本研究的基础数据，并精读属性约简研究的代表性文献 142 篇，进而总结研究热点与发展趋势。

关键词是文献内容的精髓，是对文献内容的高度凝练和总结。高频关键词在一定程度上能反映该领域研究热点和发展趋势。因此关键词词频分析对研究文章内容具有十分重要的作用，是研究某一领域热点的重要依据[4]。利用 CiteSpace 软件统计关键词词频，将词频大于 3 的 86 个关键词作为类别划分的依据。

第一类关键词涉及信息熵、条件熵、差别矩阵与正区域等，因此将其归纳为基于粗糙集的属性约简方法。第二类关键词以粒计算、知识粒度、多粒度为主，因此定义为基于粒计算的属性约简方法。第三类关键词围绕形式背景、概念格、决策形式背景等方面，故定义为基于形式概念分析的属性约简方法。第四类关键词有遗传算法、支持向量机、蚁群优化、神经网络等，主要围绕相关算法，故定义为基于智能算法的属性约简方法。

3 相关概念

3.1 粗糙集

粗糙集理论由波兰数学家 Pawlak 于 1982 年提出[5]，能够分析不确定、不完备数据，被应用于机器学习、模式识别、决策分析与知识发现等领域，其主要特点是在不改变分类能力的情况下，通过剔除冗余信息获得知识的属性约简，进而导出问题的决策规则。

设 $IS=\{U，A，V，f\}$ 为信息系统，其中，$U=\{X_1，X_2,\cdots，X_m\}$ 是对象集，$X\subseteq U$，U|A 是 U 的一个划分，X 的粗糙下、上近似分别定义如公式（1）和公式（2）所示。

$$\overline{A}X=U\{Y\in U|A：Y\subseteq X\} \quad (1)$$

$$\underline{A}X=U\{Y\in U|A：Y\cap X\neq\phi\} \quad (2)$$

相应的序对 $\langle X\underline{A},\overline{A}X\rangle$ 被称为粗糙集。

3.2 概念格

概念格[6]理论是知识表示和发现的有效工具。主要应用于认知计算、机器学习、模式识别、专家系统、决策分析、web 搜索等领域。概念格本质上描述了对象和属性之间的关联。

概念格的节点之间可以建立一种偏序关系。如果 $C_1(X_1，B_1)$，$C_2(X_2，B_2)$满足 X1 ⊆ X2 或 B2 ⊆ B1，那么 $C_1(X_1，B_1)$称为子概念，$C_2(X_2，B_2)$称为父概念，可以表示为 $C_1(X_1，B_1)\leqslant C_2(X_2，B_2)$。关系≤称为概念的偏序。对于形式语境中的所有概念，偏序集产生概念格 L(U,A,I)，对于特定的形式语境 K(U，A，I)是唯一的。

3.3 粒计算

Zadeh 首次提出粒计算（Granular Computing）的概念[7]，学界普遍认为粒计算是一种看待客观世界的世界观和方法论[8]，采用粒度思想，将复杂问题转化为简单问题的求解方法[9]，其数学描述：在空间 X 上的粒 A 可表示为空间的映射：

$$A：X\rightarrow G(X) \quad (3)$$

其中，X 表示为某空间，G 表示粒 A 的形式框架。

3.4 形式概念分析

智能医疗步研究。形式概念分析是德国数学家 Wille 提出的数据分析和规则提取方法[10]，形式背景 $K=(U，A，I)$ 其中 $U=\{X_1，X_2,\cdots，X_m\}$ 是对象集，$A=\{a_1，a_2,\cdots，a_m\}$是属性集，I 是 $U\times A$ 上的二元关系，通过概念格所展现出的概念之间的泛化与特化关系，描述对象与属性之间的依赖关系。此外，从决策信息系统 T 诱导出决策形式背景 $\prod=(U，A，I，D，J)$，其中 U，A，D 定义如公式（4）——公式（6）所示。核心概念为形式背景、形式概念与概念格。

$$U=\{X_1，X_2,\cdots，X_m\} \quad (4)$$

$$A=\{a_1(v_1^1),\cdots,\ a_1(v_{1m1}^1),\ a_2(v_1^2),\cdots,\ a_2(v_{1m2}^2),\ a_n(v_1^n),\cdots,\ a_n(v_{1mn}^n)\} \quad (5)$$

$$D=\{d(v_1^d),\ d(v_2^d),\cdots,d(v_m^d)\} \quad (6)$$

3.5 粗糙集和形式概念之间的联系

粗糙集理论与形式概念分析是两种不同的数学方法，但都是在分类基础上，从不同的侧面研究和表示数据中隐含的知识。粗糙集利用等价关系对数据表进行分类，而概念格是利用序理论对数据表进行概念分层讨论[10]。粗糙集理论是在不可分辨关系基础上进行论域的划分，该划分是一些信息粒的合集，故粗糙集方法是在单粒度空间进行的概念近似逼近，被称为单粒度粗糙集。而在概念格理论中，粒子的表述就是一个概念，包括概念的内涵与外延。

从基本组成来看，粗糙集理论是在不可分辨关系基础上，得到论域的一个划分，由上近似集和下近似集组成[11]。而粒计算由粒子、粒层与粒结构组成；形式概念分析利用序理论对数据表进行概念分层，由属性集和对象集组成。粗糙集理论通常先对连续型数据进行离散化处理，这必将造成信息损失。粒计算能够在不同粒层之间相互转化，能够高效地实现复杂问题求解。形式概念分析用概念格展现对象与属性之间的二元层次关系[12]。

4 属性约简理论方法

4.1 基于粗糙集的属性约简方法

经典粗糙集理论主要处理完备信息系统数据，学者们纷纷对经典粗糙集理论模型进行改进，将等价关系扩充为容差关系或非对称相似关系等，使其适用于不完备数据的处理，如覆盖粗糙集、概率粗糙集、变精度粗糙集、领域粗糙集、决策粗糙集等，上述模型均可以与属性约简方法相结合。经过归纳总结，典型属性约简算法包括：基于信息熵的属性约简、基于差别矩阵的属性约简和基于正区域的属性约简。

（1）基于信息熵的属性约简方法

信息熵是对事件的不确定程度的度量。假设一个总体包含 m 个随机变量Xi，i=1，2，…，m，则信息熵为：

$$H(X)=-\sum_{i=1}^{m}P(x_i)\log(x_i) \quad (7)$$

基于信息熵的属性约简的基本思路为：根据属性重要度或属性关联度，依次剔除无关属性，直到所得的属性集与原信息系统的分类能力相同为止。从信息的角度，苗夺谦等提出粗糙集中概念的信息表示，充分利用信息熵度量不确定性数据的优势，将信息熵作为启发式信息，提出高效的属性约简方法。但当条件属性较多时，时间复杂度会相应增加。Wang 等提出以条件信息熵为启发条件的约简算法，利用条件信息熵度量属性集之间依赖程度。而后，陈媛等将条件信息熵和属性重要度相结合，以决策表的相对核为起点，逐步添加引起互信息量变化大的属性，加快了决策表的运行速度。针对数据集的特异性问题，马斌斌提出基于奇异值分解熵的属性约简算法，以时间序列数据为例，与条件信息熵对比，该方法在约简结果和识别精度方面具有一定优势。 从属性区分能力角度，陶午沙等扩展基于信息熵的度量方法，提出不完备信息系统中的不确定性度量方法α信息熵，其仅针对等价关系和容差关系进行分析，缺乏对可调参数的解释；在此基础上，滕书华考虑到数据集中样本权重的不同，提出加权α信息熵的属性约简算法，融入主观偏好和先验知识，实验结果证明该算法提高了属性约简结果的分类能力。

（2）基于差别矩阵的属性约简方法

在决策统$T=(U,A,V,f)$，$U=\{X_1,\ X_2,\cdots,\ X_m\}$为对象集，$A=C\cup D$，$C=\{C_1,\ C_2,\cdots,\ C_m\}$为条件属性，$D=\{d\}$为决策属性，依据决策属性D将对象全集U划分为不同的类，即$\{X_1,\ X_2,\cdots,\ X_n\}$，则条件属性C的差别矩阵为$M(C)=|m_{ij}|_{n\times n}$定义如公式（8）所示。

$$m_{ij}=\begin{cases}c\in C,\ c(x_i)\neq c(x_j)\ \text{and}\ d(x_i)\neq d(x_j)\\ -1,\ c(x_i)=c(x_j)\ \text{and}\ d(x_i)=d(x_j)\\ \phi,\ d(x_i)=d(x_j)\end{cases} \quad (8)$$

因此，差别矩阵中元素m_{ij}能够区分对象x_i与x_j的所有属性构成的集合。当$d(x_i)=d(x_j)$时，m_{ij}的值为空集ϕ。

基于差别矩阵属性约简的基本思路为利用差别

矩阵导出区分函数，然后求解区分函数的析取范式，其中每一个析取项即为系统的一个约简。在此方法基础上，有学者提出众多改进算法，如 Felix 等提出基于二进制的差别矩阵，该差别矩阵元素由 0 和 1 组成，使存储空间减少一半；杨传健等改变存储策略，提出将垂直分解的二进制差别矩阵存储于外部介质中，仅将所需运算的二进制属性列调入内存。

（3）基于正区域的属性约简方法

基于正区域的属性约简主要是对等价类进行划分，不需要建立差别矩阵，降低了时间和空间复杂度。吴守领等通过优化终止条件，使其能够适应较大数据集的属性约简。邓大勇等提出可变正区域约简，允许正区域在一定的范围内发生变化，从而提高泛化能力。徐章艳等利用基数排序算法改进传统等价类划分算法，采用快速缩小搜索空间的思想计算属性重要度，实验证明该算法能处理大型决策表。在此基础上，葛浩等分析上述算法的局限性，以核属性为初始约简集，将重要性大的属性依次加入其中，优化等价类划分和正区域求解过程。

4.2 基于形式概念分析的属性约简方法

Ganter 等通过运用可约对象和可约属性[13]，提出了一种形式背景属性约简方法。张文修等提出基于概念格的属性约简方法，即寻找极小属性子集，使其能确定形式背景上的概念及其层次结构[14]。李进金等引入形式背景中概念格的交可约元概念并进一步提出形式背景中概念格属性约简方法[15]。在形式背景概念的基础上，张文修等提出决策形式背景概念[16]，将其定义为具有一组条件属性和一组目标属性的形式背景，由这两组属性分别形成两个概念格。随后不少学者深入研究决策形式背景的属性约简，如魏玲等提出决策形式背景的概念格属性约简理论。而后，有学者结合模糊理论，提出模糊概念格理论，但随着数据规模的增加，模糊形式背景下生成的概念数急剧增加，其相应的格结构更复杂，这使得对概念格的分析不能有效运行。

5 属性约简方法应用趋势

5.1 动态属性约简算法

现实世界存在着大量不断变化的数据，如果每获得一批数据，都对数据集重新进行属性约简计算，必将造成不必要资源损耗。针对动态数据中数据对象增加、属性增加、属性值变化三种情形，产生了很多算法。

（1）数据对象动态增加

针对此类数据，很多学者提出基于 Skowron 区分矩阵的增量算法，但不能处理不协调决策表。有学者认为在更新差别矩阵时，仅须插入某一行及某一列，或删除某一行并修改相应的列，Liang 等通过分析增加一组数据后样本集上数据分布的变化，提出使用信息熵的增量机制来确定属性重要度的方法。当同时增加多个数据对象时，Shu 等提出基于正区域的增量属性约简方法，而 Jing 等提出基于知识粒度的增量式约简方法。当新增数据对象与原约简集完全矛盾时，申雪芬找到区分新增数据对象和矛盾对象的属性，进而转化为新增属性问题。

（2）数据维度动态增长

许多真实数据集不仅对象在增加，而且属性维度也在动态变化。以医疗诊断决策系统为例，已经存在不同的临床特征，如头疼、温度、血压等数据信息，新的临床特征会逐步被添加，即属性维度会增加。针对此类数据，维度增量属性约简算法被提出。Shu 针对不完备决策系统中添加和删除属性集时，提出正区域的属性约简算法。Wang 等提出基于信息熵的维度增量属性约简算法。景运革探讨当多个对象的属性值发生变化后，引入知识粒度增量机制。

（3）属性值变化

Wang 等分析对象的属性值发生变化时，在互补熵、组合熵和香农熵的基础上，提出一种属性值动态变化增量属性约简算法。王磊等针对属性值发生粗化、细化变化情况下，设计概念近似集增量式更新的矩阵算法。同样，针对此种情况，季晓岚等提出优势关系下决策信息系统近似集增量更新算法。

5.2 大数据背景下属性约简方法

（1）数据集分解方法

传统的属性约简方法将整个数据集一次性装入内存，很难处理大规模数据。Kusiak 采取数据集分的方法，减少每次处理的数据量，提高属性约简算法计算效率。

（2）并行属性约简方法

传统的并行属性约简利用任务并行计算属性约简，因此，有学者融入并行计算的思想，将属性约简任务分配到多个中央处理器中同时进行，从而提高属性约简的效率。此外，Deng 等提出并行约简思想，将大规模数据分解为多个子决策表，分别对各个子决策表计算正区域个数，选择一个最优候选属性，重复此过程，直到获取一个约简；而 Liang 等分别计算子决策表上的每一个约简，然后融合各个约简，得到最

终约简结果。随着大数据平台的迅速发展，并行计算编程模式 MapReduce 逐渐成熟，大数据环境下属性约简算法研究不断涌现。学者们纷纷使用 Hadoop 平台和 MapReduce 分布式计算框架，对粗糙集属性约简在云环境下进行分析实现。

6 总结

属性约简是数据挖掘、知识发现等领域中的重要研究议题，研究者们不断努力改进算法，但当前属性约简研究的数据源多以少量数据进行实验，属性约简结果评价指标和方式单一，基本上都是先约简属性，然后以分类结果的 F1 值或时间和空间复杂度作为评价指标。随着数据呈现海量、高维及动态等特征，属性约简方法遇到了前所未有的挑战。

目前，大数据环境下属性约简还不够成熟，例如对于海量数据、混合型数据、数据缺失及不协调问题，尚缺乏高效的启发式算法，需要关注算法中的数据结构以及内存管理问题，如何根据实际情况融合多种智能算法，与其他处理不确定知识的方法相结合，提高属性约简算法的效率，是今后主要的研究方向之一。

随着大数据技术和分布式数据存储技术的成熟，数据集分解方法日趋完善，但各数据子集之间存在差异，构建分布式环境进行数据子集的并行处理，采取合适的属性加权方法，提出分布式属性约简算法，是有待于进一步研究的问题。

粗糙集属性约简通常先对连续型数据进行离散化处理，必将造成信息损失。模糊集理论则关注信息系统中知识的模糊性。因此可将粗糙集与模糊相结合对连续型数据表示的对象进行聚类划分，将属性的模糊性转化为对象的模糊性。粗糙集理论定义的分类边界过于简单，产生的决策规则不太稳定，而且分类精确性不高。基于神经网络的属性约简因为属性众多，数据规模庞大，存在网络结构复杂、约简速度慢等问题，两者结合可以很好地弥补各自的缺点。

参考文献

[1] Lin K C, Zhang K Y, Huang Y H, et al. Feature Selection Based on an Improved Cat Swarm Optimization Algorithm for Big Data Classification[J]. Journal of Supercomputing, 2016, 72(8): 3210-3221.

[2] Yao Y, Zhao Y. Attribute Reduction in Decision-Theoretic Rough Set Models[J].Information Sciences,2008,178(17):3356-3373.

[3] Wong S K M, Ziarko W. On Optimal Decision Rules in Decision Tables[J]. Bulletin of the Polish Academy of Sciences Mathematics,1985,33(11-12):693-696.

[4] 谭章禄,彭胜男,王兆刚 .基于聚类分析的国内文本挖掘热点与 趋势研究[J].情报学报,2019,38(6):578-585.（ Tan Zhanglu, Peng Shengnan, Wang Zhaogang. Research on Hotspots and Trends of Domestic Text Mining Based on Cluster Analysis[J].Journal of the China Society for Scientific and Technical Information, 2019, 38(6):578-585.）

[5] Abdolrazzagh-Nezhad M. Enhanced Cultural Algorithm to Solve Multi-objective Attribute Reduction Based on Rough Set Theory [J]. Mathematics and Computers in Simulation, 2020, 170: 332-350.

[6] Yao Y. Three-Way Decision: An Interpretation of Rules in Rough Set Theory[C]// Proceedings of the 2009 International Conference on Rough Sets and Knowledge Technology. Springer Berlin Heidelberg, 2009:642-649.

[7] Zadeh L A. Toward a Theory of Fuzzy Information Granulation and Its Centrality in Human Reasoning and Fuzzy Logic[J]. Fuzzy Sets & Systems, 1997, 90:111-127.

[8] Li J, Huang C, Qi J, et al. Three-Way Cognitive Concept Learning via Multi-Granularity[J]. Information Sciences, 2017, 378: 244-263.

[9] 姚一豫, 祁建军, 魏玲. 基于三支决策的形式概念分析、粗糙集与粒计算[J]. 西北大学学报:自然科学版, 2018, 48(4):477-487. (Yao Yiyu, Qi Jianjun, Wei Ling. Formal Concept Analysis, Rough Set Analysis and Granular Computing Based on Three Way Decisions[J]. Journal of Northwest University: Natural Science Edition, 2018, 48(4):477-487.)

[10] Ganter B, Godin R. Formal Concept Analysis[M].Springer Berlin Heidelberg, 1999.

[11] 王虹,张文修.形式概念分析与粗糙集的比较研[J].计算机工 程,2006,32(8):42-44.（Wang Hong, Zhang Wenxiu. Comparative Study on Formal Concept Analysis and Rough Set Theory[J]. Computer Engineering, 2006, 32(8): 42-44.）

[12] Pawlak Z. Rough Sets: Theoretical Aspects of Reasoning About Data[M]. London: Kluwer Academic Publishers,1991.

[13] 李进金,张燕兰,吴伟志,等.形式背景与协调决策形式背景属性约简与概念格生成[J]. 计算机学报, 2014,37(8):1768-1774.（ Li Jinjin, Zhang Yanlan, Wu Weizhi, et al. Attribute Reduction for Formal Context and Consistent Decision Formal Context and Concept Lattice Generation[J]. Chinese Journal of Computers, 2014,37(8):1768-1774.）

[14] Chen D, Wang C, Hu Q. A New Approach to Attribute Reduction of Consistent and Inconsistent Covering Decision Systems with Covering Rough Sets[J].Information Sciences, 2007, 177(17): 3500-3518.

[15] Katzberg J D, Ziarko W. Variable Precision Extension of Rough Sets [J]. Fundamenta Informaticae, 1996,27(2-3):155-168.

[16] Lin T Y. Granular Computing on Binary Relations I: Data Mining and Neighborhood Systems[J]. Rough Sets in Knowledge Discovery,1998, 1: 107-121.