《智能信息处理》课程作业

基于粒计算的多粒度数据分析方法综述

米富横

作业	分数[20]
得分	

2021 年 11 月 29 日

基于粒计算的多粒度数据分析方法综述

米富横

（大连海事大学　信息科学与技术学院　大连　116026）

摘　要　多粒度数据是一种特殊的、有用的数据类型，它通过对论域（研究对象的集合）采用不同的粒化方式使得数据能够在多个粒度空间中进行呈现，在此基础上可以开展数据的多层次知识发现研究。商空间理论、序贯三支决策、多粒度粗糙集、多尺度数据分析模型和多粒度形式概念分析是几种常见的、有效的多粒度数据分析方法，已受到人们的广泛关注。本文对基于粒计算的多粒度数据分析研究工作进行综述，给出每一类多粒度数据分析方法的理论框架、基本概念以及主要研究思想，并指出多粒度数据分析研究中存在的若干问题，为该领域的后续研究提供理论参考。

关键词　粒计算；多粒度粗糙集；多粒度形式概念分析

Review of Multi-granularity Data Analysis Methods Based on Granular Computing

Mi Fuheng

(Department of Information Science Technology, Dalian Maritime University, Dalian　116026)

Abstract　Multi-granularity data is a special useful type of data which is able to show data in different granularity spaces by using different granularity forms of a universe of discourse（i. e. a set of research objects）, and then multi-level knowledge discovery can be studied based on multi-granularity data. As is well-known, quotient space theory, sequential three-way decision, multi-granulation rough set,multi-scale data analysis model and multi-granularity formal concept analysis are several common and effective multi-granularity data analysis methods, and they have attracted more and more people's attention. This paper reviews the existing work on multi-granularity data analysis in granular computing, gives theoretical frameworks, basic notions and main research ideas for each kind of multi-granularity data analysis methods, and points out some problems for the further study of multi-granularity data analysis. The obtained results can provide a theoretical reference for future research of this field.

Keywords　granular computing; multi-granulation rough set; multi-granularity formal concept analysis

1 引 言

多粒度数据分析是大数据研究领域中的重要课题，它基于多粒度思想对数据进行多角度、深层次的分析与处理，以解决现实中特定复杂数据的知识发现与表示问题[1]。多粒度数据通常是对数据进行不同粒化得到的多个侧面认识，它的显著特点是数据可以在多个粒度空间进行呈现。多粒度数据广泛存在于人们生产生活中，比如考试成绩有百分制也有等级制，电子地图能够显示到市县一级也能够进一步细化到乡镇一级甚至村落一级，论文可以按影响因子排名也可以按分区明确其大致的等级。也就是，人们会根据自己的实际决定在哪个层面进行数据分析与处理比较合适，这也是人们解决问题时比较擅长的方面。实际上，出现这种现象是由于数据在不同粒度层面上进行显示的

代价和效果存在差异导致的。

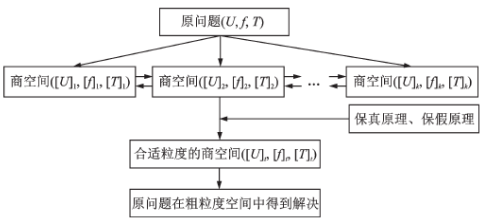
对论域进行粒化通常能够带来计算上的便利以及结果的表示会更加简洁等诸多好处，而本质则是提高解决问题所需的解空间的维度，使得解空间的搜索范围相对变小（即粒度变大使得总的粒度个数变少），这种粒度转化思维尤其适用于完成大规模数据分析任务。

本文不讨论这些粒计算方法之间存在的区别与联系，而是以多粒度数据分析为研究对象，归纳总结当前粒计算领域中已开展的多粒度数据分析工作。

2 基于商空间理论的多粒度数据分析

假设用三元组 (U，f，T)描述一个复杂问题，其中 U 代表问题的论域，f(•)代表论域的属性，T 代表论域的结构。商空间理论[6]主要研究如下形式的问题：在给定关系 R（可以多个关系）的情况下，将论域 U 自然投影到商集 [U]，同时伴随着论域的属性 f 和结构 T 也自然投影到商集 [f] 与 [T]，那么原问题(U，f，T) 与商空间 ([U]，[f]，[T]) 之间关于问题的解存在哪些关系？当给定多个关系 Ri(i=1，2，…，k)时，商空间 ([U]i，[f]i，[T]i)之间关于问题的解又存在什么关系？商空间理论的大致框架如图 1 所示，它的核心思想是探讨一个复杂问题如何通过粒化分析与运算能够在粗粒度世界中得到解决，以降低求解问题的复杂性。需要指出的是，保真原理和保假原理为粗细粒度世界中的解进行相互推理起到了关键作用[2]。

图 1 空间商理论构架



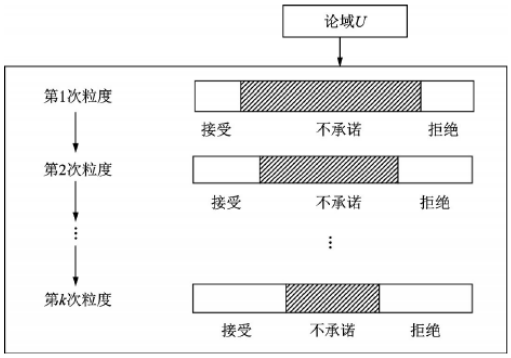
自商空间理论提出以来，已受到粒计算领域的专家和学者的广泛关注与研究。文献通过模糊等价关系将精确粒度下的商空间理论推广到模糊粒度环境中，在模糊粒度计算、粗糙集和商空间三者之间搭建起联系的桥梁，为交叉融合研究做出了贡献。

商空间理论是一种特色比较鲜明的多粒度数据分析方法，它在粒化论域的同时也对属性和结构进行粒化，某种程度上体现了特定的拓扑结构思想。

3 基于序贯三支决策的多粒度数据分析

序贯三支决策[3]是三支决策理论中用于处理多粒度数据的一种主要分析方法，它的核心思想如图 2 所示，也就是通过某种方式反复对论域 U 进行三分粒化（通常借助于 0 到 1 之间的 2 个阈值参数 α，β（α < β）来完成，即满足约束条件大于等于 β 的对象分成一类，小于等于 α 的对象归为另一类，介于它们之间的对象又是一类），在此基础上讨论三分类问题，通常称这三个类分别为接受域、拒绝域和不承诺部分。实际上，不承诺意味着不确定、暂时不采取措施的意思，接受或拒绝则表明已明确其决策结果。一般地，为了解决具体的实际问题，对论域 U 进行反复粒化是一个不确定域（不承诺部分）逐渐减少的过程，因为序贯三支决策的目的并非不进行决策，而是通过暂时延缓决策以减少误判等行为而付出的额外代价。众所周知，随着时间的推移，如果信息能够得到不断完善或补充，那么不确定域最终有望变成空集，此时三分类问题就退化为二分类问题，从而做出明确的误分类代价或测试代价较少的决策。当然，在最坏的情况下，很可能最终仍有一部分对象无法进行明确的决策，此时不得已只能随机将其划分到接受域或拒绝域。需要指出的是，到底是何种原因导致人们对论域 U 进行反复粒化，这取决于拟解决问题的具体需求。

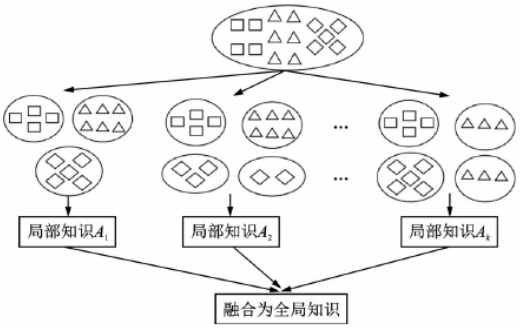
图 2 序贯三支决策



4 基于多粒度粗糙集的多粒度数据分析

乐观与悲观多粒度粗糙集是信息融合（多约束问题）的两种极端情况，即要么满足一个约束条件即可，要么所有的约束条件都必须满足[4]。当然，推广到部分约束条件成立的多粒度粗糙集也不是难事，它是一种折中的信息融合模式，但是到底哪些项应处理成乐观约束条件而另外的当作悲观约束条件，则需要根据具体问题进行具体分析。更一般地，约束条件的权重有时不一定相等，也就是研究代价敏感多粒度粗糙集也是有必要的。此外，将上述模型中的属性子集替换成近似空间，可以得到通用的多粒度粗糙集模型，详见文献。根据上述讨论，多粒度粗糙集的核心思想如图 3 所示，其中融合策略可根据具体问题的研究背景灵活选取。

图 3 多粒度粗糙集的核心思想

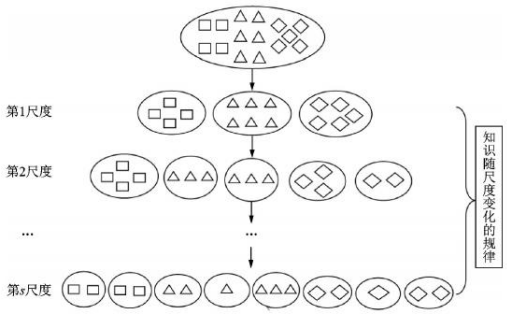


多粒度粗糙集是信息融合的一种有效方法，根据诱导信息的关系不同以及应用的差异已提出多种数学模型，它的特点是依据实际需要对各种粒度进行交叉融合运算，以解决多粒度数据分析问题。

5 基于多尺度数据分析模型的多粒度数据分析

Wu-Leung 模型[5]是多尺度数据分析的原型方法，它的核心思想如图 4 所示，图中每个尺度下的数据对应着论域 U 的一次粒化结果，且这些粒化结果之间能够形成粒度粗细关系，即信息呈现出从粗粒度到细粒度或从细粒度到粗粒度的规律。多尺度数据可简单理解为包含多尺度属性的信息系统，因此简称为多尺度信息系统，通常记为二元组 (U，C)，其中 C ={akj|j = 1，2，…，m；k = 1，2，…，s}，m 为属性个数，s 为尺度个数。也就是，每个对象在同一个属性下会有多个取值，实际上它是同一个值在不同粒度空间中的不同表现形式，本质上仍为一个取值。

图 4 Wu-Leung 模型的核心思想



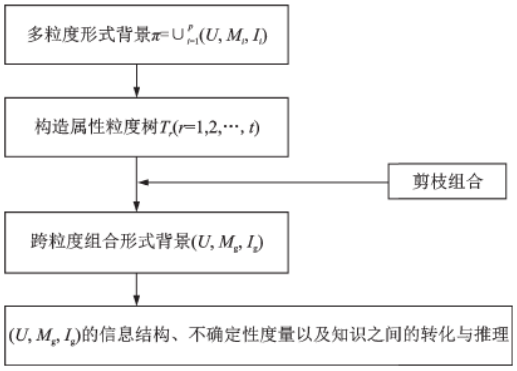
继 Wu-Leung 模型之后，一个有意思的问题是，在满足给定要求的情况下如何选取合适的尺度？于是，Wu 和 Leung 又提出了最优尺度选择问题，以寻找知识随尺度变化过程中出现的一个临界理想状态。总之，Wu-Leung 模型针对多尺度数据研究知识随尺度变化的规律以及知识在哪个尺度下进行分析最为合适等问题，截至目前已取得系列成果，具有广阔的应用前景和发展潜力。

6 基于多粒度形式概念分析的多粒度数据分析

多粒度形式概念分析从多粒度形式背景π = ∪i = 1p(U，Mi，Ii)（多个单粒度形式背景的并置得到）出发，然后构

造属性粒度树 Tr(r =1，2，…，t)（同类属性信息之间的层次结构），再对属性粒度树进行剪枝（即粒度树节点之间的一种信息组合方式）形成跨粒度组合形式背景(U，Mg，Ig)，在此基础上讨论所得的跨粒度组合数据的拓扑结构、不确定性度量以及知识之间的转化与推理等，具体如图 5 所示。类似于多尺度信息系统，多粒度形式背景的信息之间也是满足粒度粗细关系的[6]。

图 5 多粒度形式概念分析的核心思想



在机器学习领域中，也存在一些多粒度数据分析方法，其中多粒度学习就是基于粒计算而提出的有效方法之一，其中分类训练包括 2 大类：（1）标签结构信息的训练，也就是多粒度分类问题，即分类任务与粒度密切相关，有些样本在粗粒度中进行分类即可，不宜在细粒度中分类（容易出现误分类现象），而有些样本则需要具体到最细的粒度层面；（2）传统的分类训练，即训练一个单粒度分类器。一般地，多粒度学习的特点是通过对数据粒度进行不断调整，以得到研究对象比较满意的分类结果（有时还包括分类间的结构信息），以帮助分析数据和解决复杂问题。

7 结 论

本文对基于商空间理论、序贯三支决策、多粒度粗糙集、多尺度数据分析模型和多粒度形式概念分析等的多粒度数据分析研究工作进行了综述，指出了每一类多粒度数据分析方法的核心思想、基本问题以及主要研究思想。

除了本文介绍的多粒度数据分析方法之外，基于粒计算的类似研究还有很多，因此本文只是对该研究领域做了一个大致的综述，并未窥其全貌，也没有将该问题的所有重要工作都归纳总结到位。相信今后会涌现出更多、更好的多粒度数据分析方法，进一步促进该研究领域的快速发展，这方面取得的研究成果将成为粒计算在数据分析应用中的标志性工作。

参 考 文 献

[1] 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: 信息科学, 2015, 45(11): 1355-1369.

[2] WANG Guoyin, XU Ji. Granular computing with multiple granular layers for brain big data processing [J]. Brain Informatics,2014, 1(1/2/3/4): 1-10.

[3] 徐计,王国胤,于洪 . 基于粒计算的大数据处理[J]. 计算机学报, 2015, 38(8): 1497-1517.

[4] 苗夺谦,张清华,钱宇华,等 . 从人类智能到机器实现模型——粒计算理论与方法[J]. 智能系统学报,2016, 11(6): 743-757.

[5] WU Weizhi, LEUNG Y. Theory and applications of granular labelled partitions in multi-scale decision tables[J]. Information Sciences, 2011, 181(18): 3878-3897.

[6] 张钹,张铃 . 问题求解理论及应用[M]. 北京: 清华大学出版社, 1990.