

/// 粒计算研究丛书 ///

云模型 与 粒计算

王国胤 李德毅 姚一豫 梁吉业 等著
苗夺谦 张燕平 张清华



科学出版社

内 容 简 介

云模型是研究定性概念与定量数值之间相互转换的不确定性认知模型。粒计算是当前计算智能研究领域模拟人类思维和解决复杂问题的新方法。它覆盖了所有有关粒度的理论、方法和技术,是研究复杂问题求解、海量数据挖掘和模糊信息处理等问题的有力工具。本书介绍云模型与粒计算交叉研究的最新进展,由国内外相关领域的华人学者撰文 14 章,内容涉及云模型、高斯云的数学性质、云模型与相近概念的关系、区间集、区间值信息系统的粒计算模型与方法、多粒度粗糙集、粒计算模型的特性分析与比较、云计算环境下层次粗糙集模型约简算法、基于粒计算的聚类分析、并行约简与 F -粗糙集、单调性分类学习、不确定性研究中若干问题的探讨、基于云模型的文本分类应用、数据挖掘算法的云实现。

本书可供计算机、自动化等相关专业的研究人员、教师、研究生、高年级本科生和工程技术人员参考。

图书在版编目(CIP)数据

云模型与粒计算/王国胤等著. —北京: 科学出版社, 2012
(粒计算研究丛书)
ISBN 978-7-03-035064-0

I. ①云… II. ①王… III. ①人工智能-算法-研究 IV. ①TP18

中国版本图书馆 CIP 数据核字(2012)第 148151 号

责任编辑: 任 静 / 责任校对: 宋玲玲
责任印制: 张 倩 / 封面设计: 速底书装

科学出版社 出版

北京东黄城根北街 16 号
邮政编码: 100717

<http://www.sciencep.com>

源海印刷有限责任公司印刷

科学出版社发行 各地新华书店经销

*

2012 年 7 月第 一 版 开本: B5(720×1000)

2012 年 7 月第一次印刷 印张: 21

字数: 406 000

定价: 75.00 元

(如有印装质量问题,我社负责调换)

《粒计算研究丛书》编委会

名誉主编：李德毅 张 钹

主 编：苗夺谦 王国胤 姚一豫

副主编：梁吉业 吴伟志 张燕平

委 员：(按拼音排序)

陈德刚 代建华 高 阳 胡清华

胡学钢 黄 兵 李德玉 李凡长

李进金 李天瑞 刘贵龙 刘 清

米据生 史开泉 史忠植 王飞跃

王 珏 王熙照 徐久成 杨 明

姚静涛 叶东毅 于 剑 曾黄麟

张 铃 张文修 周献忠 祝 峰

秘 书：王睿智 张清华

丛书序

粒计算是一个新兴的、多学科交叉的研究领域。它既融入了经典的智慧,也包括了信息时代的创新。通过十多年的研究,粒计算逐渐形成了自己的哲学、理论、方法和工具,并产生了粒思维、粒逻辑、粒推理、粒分析、粒处理、粒问题求解等诸多研究课题。值得骄傲的是,中国科学工作者为粒计算研究发挥了奠基性的作用,并引导了粒计算研究的发展趋势。

在过去几年里,科学出版社出版了一系列具有广泛影响的粒计算著作,包括《粒计算:过去、现在与展望》、《商空间与粒计算——结构化问题求解理论与方法》、《不确定性与粒计算》等。为了更系统、全面地介绍粒计算的最新研究成果,推动粒计算研究的发展,科学出版社推出了《粒计算研究丛书》。本丛书的基本编辑方式为:以粒计算为中心,每年选择该领域的一个突出热点为主题,邀请国内外粒计算和该主题方面的知名专家、学者就此主题撰文,来介绍近期相关研究成果及对未来的展望。此外,其他相关研究者对该主题撰写的稿件,经丛书编委会评审通过后,也可以列入该系列丛书。本丛书与每年的粒计算研讨会建立长期合作关系,丛书的作者将捐献稿费购书,赠给研讨会的参会者。

中国有句老话,“星星之火,可以燎原”,还有句谚语,“众人拾柴火焰高”。《粒计算研究丛书》就是基于这样的理念和信念出版发行的。粒计算还处于婴儿时期,是星星之火,在我们每个人的爱心呵护下,一定能够燃烧成燎原大火。粒计算的成长,要靠大家不断地提供营养,靠大家的集体智慧,靠每一个人的独特贡献。这套丛书为大家提供了一个平台,让我们可以相互探讨和交流,共同创新和建树,推广粒计算的研究与发展。本丛书受益于粒计算研究每一位同仁的热心参与,也必将服务于从事粒计算研究的每一位科学工作者、老师和同学。

《粒计算研究丛书》的出版得到了众多学者的支持和鼓励,同时也得到了科学出版社的大力帮助。没有这些支持,也就没有本丛书。我们衷心地感谢所有给予我们支持和帮助的朋友们!

《粒计算研究丛书》编委会
2012年7月

前 言

21 世纪互联网和云计算广泛应用,社会计算和智能计算快速发展,分享、交互和群体智能改变了人类的生活方式,人们迫切期望着智慧地球时代的来临,不确定性人工智能的新时代已经到来。认知的不确定性,必然导致不确定性人工智能的研究。研究不确定性知识的表示、处理和模拟,寻找并且形式化地表示不确定性知识中的规律性,让机器模拟人类认识客观世界和人类自身的认知过程,使机器具有不确定性智能,成为人工智能学家的一项重要任务。2011 年的图灵奖授给了对不确定性推理做出杰出贡献的朱地亚·佩尔(Judea Pearl)教授就是一个证明。用自然语言的方法把握量的不确定性,比数学表达更真实、更具有普适性。那么,如何表示用自然语言表述的定性知识呢?如何反映自然语言中的不确定性,尤其是模糊性和随机性呢?怎样实现定性和定量知识之间的相互转换?怎样体现语言思考中的软推理能力呢?云模型(cloud model)作为一种研究定性 with 定量相互转换的不确定性认知模型,可以实现定性概念与定量数值之间的双向转换,对于这些问题进行了有益的探索。

粒计算(granular computing)是当前计算智能研究领域模拟人类思维和解决复杂问题的新方法。它覆盖了所有有关粒度的理论、方法和技术,是研究复杂问题求解、大数据挖掘和不确定性信息处理等问题的有力工具。粒计算方法的研究是近几年人工智能领域中崛起的一个新方向,它从实际出发,用尽力而为的满意解替代精确解,其主要思想是在不同的粒度层次上进行问题求解,在很大程度上体现了人类问题求解过程中的智能。粒计算从提出到现在已有 30 多年,近年来受到众多研究者的广泛关注,已经成为日益受到学术界重视的一个新研究领域。

随着粒计算研究工作的不断深入,人们从不同的角度研究得到不同的粒计算理论模型,主要有模糊集(词计算)理论模型、粗糙集理论模型、商空间理论模型和云模型等。我国学者开展了以国际、国内学术研讨会议和暑期研讨会等多种形式的粒计算学术交流与合作,相继出版了一系列著作,如 2007 年张钹、张铃的《问题求解理论及应用——商空间粒度计算理论及应用(第 2 版)》(清华大学出版社),2008 年由 13 位海内外华人学者合著的《粒计算:过去、现在与展望》(科学出版社)。结合粒计算专题,国内举办了两次专题研讨会,即 2010 年 7 月在安徽大学举办了商空间与粒计算的专题研讨会,出版了《商空间与粒计算——结构化问题求解理论与方法》(科学出版社),2011 年 8 月在同济大学举办了不确定性与粒计算的专题研讨会,出版了《不确定性与粒计算》(科学出版社)。另外,“中国粒计算学术研讨会”、“IEEE International Con-

ference on Granular Computing”、“Advances in Granular Computing”等国内外有关学术活动的开展,也极大地促进了粒计算理论及其应用的发展。

今年,在中国人工智能学会粗糙集与软计算专业委员会和国际粒计算系列研讨班指导委员会的共同组织下,粒计算研究领域的学者又汇聚北京,共同发起并组织2012年“云模型与粒计算暑期学术研讨会”。本次会议旨在将云模型与粒计算的研究相结合,探讨云模型与粒计算的关系,探索基于云模型的粒思维、粒推理以及粒化等不确定性人工智能方法,以获取有效的不确定性问题求解的理论与方法。期望借此会议的举办及本书的出版,在粒计算领域开辟一个新的研究方向。

本书由国内外云模型和粒计算研究领域的20余位华人学者合作撰写。书中介绍了各位作者及其研究团队近年来在云模型和粒计算研究方面取得的最新研究成果,及时总结了相关研究进展。本书除作为2012年暑期研讨会的教材外,还可为人工智能领域的专家或研究人员从事相关研究提供参考。

全书组织结构如下:第1章双向认知计算模型——云模型,由王国胤、许昌林、张清华撰写;第2章高斯云的数学性质,由刘玉超、李德毅、刘禹撰写;第3章云模型与相近概念的关系,由胡宝清撰写;第4章区间集,由姚一豫撰写;第5章区间值信息系统的粒计算模型与方法,由代建华、王文涛撰写;第6章多粒度粗糙集,由钱宇华、梁吉业撰写;第7章粒计算模型的特性分析与比较,由张铃、张燕平撰写;第8章云计算环境下层次粗糙集模型约简算法,由苗夺谦、钱进撰写;第9章基于粒计算的聚类分析,由丁世飞等撰写;第10章并行约简与 F -粗糙集,由邓大勇、陈林撰写;第11章单调性分类学习,由胡清华、潘巍巍撰写;第12章不确定性研究中若干问题的探讨,由胡宝清撰写;第13章基于云模型的文本分类应用,由何中市、代劲撰写;第14章数据挖掘算法的云实现,由刘震、周涛撰写。

由于作者水平有限,加之时间仓促,书中不妥之处在所难免,恳请读者批评指正。

本书的出版得到了国家自然科学基金(No. 61073146)、重庆邮电大学出版基金等项目的资助,在此一并致谢。

目 录

前言

| | |
|----------------------------------|----|
| 第 1 章 双向认知计算模型——云模型 | 1 |
| 1.1 引言 | 2 |
| 1.2 正态分布与正态隶属函数 | 4 |
| 1.3 云模型 | 6 |
| 1.3.1 云模型的定义 | 6 |
| 1.3.2 云模型的数字特征 | 7 |
| 1.3.3 正态云模型的递归定义及其数学性质 | 8 |
| 1.3.4 云发生器 | 8 |
| 1.3.5 双向认知计算模型 | 12 |
| 1.4 本章小结 | 20 |
| 参考文献 | 21 |
| 第 2 章 高斯云的数学性质 | 23 |
| 2.1 高斯云分布 | 23 |
| 2.2 高斯云的数学性质 | 25 |
| 2.2.1 高斯云的数字特征 | 25 |
| 2.2.2 高斯云的期望曲线 | 27 |
| 2.2.3 高斯云的雾化特性 | 28 |
| 2.2.4 高阶高斯云的数字特征 | 30 |
| 2.3 高斯云的参数对峰度的影响分析 | 32 |
| 2.4 高斯云的幂律特性实验 | 35 |
| 2.5 本章小结 | 39 |
| 参考文献 | 39 |
| 第 3 章 云模型与相近概念的关系 | 40 |
| 3.1 二型 Fuzzy 集 | 40 |
| 3.1.1 二型 Fuzzy 集的定义 | 40 |
| 3.1.2 二型 Fuzzy 集的计算 | 42 |
| 3.1.3 区间值 Fuzzy 集 | 44 |
| 3.1.4 Gaussian 二型 Fuzzy 集 | 46 |

| | | |
|--------------|--------------------------------|-----------|
| 3.1.5 | 二型 Fuzzy 集的嵌入区间值 Fuzzy 集 | 47 |
| 3.1.6 | m 型 Fuzzy 集与 Genuine 集 | 49 |
| 3.1.7 | 区间集与阴影集 | 50 |
| 3.2 | 直觉 Fuzzy 集 | 52 |
| 3.2.1 | 直觉 Fuzzy 集的定义 | 52 |
| 3.2.2 | 直觉 Fuzzy 集的运算 | 53 |
| 3.2.3 | 区间值直觉 Fuzzy 集 | 53 |
| 3.3 | Neumaier 云 | 55 |
| 3.3.1 | Neumaier 云的定义 | 55 |
| 3.3.2 | 离散云 | 58 |
| 3.3.3 | 连续云与潜云 | 58 |
| 3.4 | Fuzzy 概率集 | 59 |
| 3.4.1 | 随机集 | 59 |
| 3.4.2 | Fuzzy 概率集 | 59 |
| 3.4.3 | Bifuzzy 概率集和区间值概率集 | 61 |
| 3.5 | Soft 集 | 61 |
| 3.5.1 | Soft 集的定义 | 61 |
| 3.5.2 | Soft 集的运算 | 63 |
| 3.6 | 云模型 | 64 |
| 3.6.1 | 云模型的定义 | 64 |
| 3.6.2 | 云模型算法 | 64 |
| 3.6.3 | 正态云生成的区间值 Fuzzy 集 | 65 |
| 3.7 | 云集 | 67 |
| 3.7.1 | 各类集合的关系 | 67 |
| 3.7.2 | 云集 | 69 |
| 3.8 | 本章小结 | 69 |
| | 参考文献 | 70 |
| 第 4 章 | 区间集 | 74 |
| 4.1 | 引言 | 74 |
| 4.2 | 不精确概念及其表示 | 75 |
| 4.3 | 区间集 | 76 |
| 4.3.1 | 区间集与部分已知概念 | 76 |
| 4.3.2 | 区间集与概念近似 | 77 |
| 4.4 | 区间集代数 | 78 |

| | | |
|--------------|-------------------------|------------|
| 4.4.1 | 幂代数 | 78 |
| 4.4.2 | 区间集运算 | 78 |
| 4.4.3 | 基于包含序的区间集代数 | 80 |
| 4.4.4 | 基于知识序的区间集代数 | 81 |
| 4.5 | 基于不完备信息表的区间集构造方法 | 81 |
| 4.6 | 区间集与其他理论的联系 | 85 |
| 4.6.1 | 区间集与 Kleene 三值逻辑 | 85 |
| 4.6.2 | 区间集与粗糙集 | 86 |
| 4.6.3 | 区间集与三支决策 | 86 |
| 4.6.4 | 区间集、模糊集和云模型 | 87 |
| 4.7 | 本章小结 | 87 |
| | 参考文献 | 87 |
| 第 5 章 | 区间值信息系统的粒计算模型与方法 | 94 |
| 5.1 | 引言 | 94 |
| 5.2 | 基础概念 | 95 |
| 5.2.1 | 不可分辨关系和近似域 | 95 |
| 5.2.2 | 决策系统中的不确定性度量 | 95 |
| 5.3 | 区间值决策系统的不确定性度量 | 96 |
| 5.3.1 | 区间值的相似关系 | 96 |
| 5.3.2 | 相似类和决策类 | 97 |
| 5.3.3 | θ -条件熵 | 99 |
| 5.3.4 | 不确定性度量和 θ -粗糙决策熵 | 101 |
| 5.4 | 实验 | 103 |
| 5.5 | 本章小结 | 107 |
| | 参考文献 | 107 |
| 第 6 章 | 多粒度粗糙集 | 110 |
| 6.1 | 问题描述 | 110 |
| 6.2 | 乐观多粒度粗糙集 | 112 |
| 6.2.1 | Pawlak 粗糙集理论 | 112 |
| 6.2.2 | 乐观粗糙近似 | 113 |
| 6.2.3 | 多粒度粗糙集中的几个度量 | 122 |
| 6.2.4 | 特征选择 | 125 |
| 6.3 | 悲观多粒度粗糙集 | 128 |
| 6.3.1 | 悲观粗糙近似 | 128 |

| | | |
|------------|--------------------------------|------------|
| 6.3.2 | 粗糙成员函数 | 132 |
| 6.3.3 | 多粒度粗糙集中的规则 | 133 |
| 6.4 | 本章小结 | 134 |
| | 参考文献 | 135 |
| 第7章 | 粒计算模型的特性分析与比较 | 137 |
| 7.1 | 引言 | 137 |
| 7.2 | 不确定性表示方法 | 141 |
| 7.2.1 | 隶属度的方法 | 141 |
| 7.2.2 | 粗糙集的表示方法 | 143 |
| 7.2.3 | 商空间的表示方法 | 144 |
| 7.3 | 粒计算表示不确定性方法之间的关系 | 145 |
| 7.3.1 | 隶属度函数表示方法与商空间链表示方法的关系 | 145 |
| 7.3.2 | 粗糙集的表示方法与商空间链表示方法之间的关系 | 146 |
| 7.3.3 | 云模型与二型模糊之间的关系 | 148 |
| 7.4 | 问题求解方法的比较 | 149 |
| 7.5 | 本章小结 | 153 |
| | 参考文献 | 154 |
| 第8章 | 云计算环境下层次粗糙集模型约简算法 | 156 |
| 8.1 | 层次粗糙集模型 | 157 |
| 8.1.1 | 引言 | 157 |
| 8.1.2 | 概念层次 | 158 |
| 8.1.3 | 基于云模型的概念提取及概念提升 | 159 |
| 8.1.4 | 层次粗糙集模型 | 162 |
| 8.2 | 云计算技术 | 167 |
| 8.2.1 | 云计算介绍 | 167 |
| 8.2.2 | MapReduce 技术 | 167 |
| 8.3 | 云计算环境下层次粗糙集模型约简算法 | 168 |
| 8.3.1 | 云计算环境下知识约简算法中的并行性分析 | 168 |
| 8.3.2 | 云计算环境下计算层次编码决策表算法 | 169 |
| 8.3.3 | 云计算环境下层次粗糙集模型约简算法的研究 | 170 |
| 8.4 | 实验与分析 | 174 |
| 8.4.1 | 理论分析 | 174 |
| 8.4.2 | 实验结果 | 175 |
| 8.4.3 | 实验分析 | 175 |

| | |
|--|------------|
| 8.5 本章小结 | 178 |
| 参考文献 | 178 |
| 第9章 基于粒计算的聚类分析 | 182 |
| 9.1 引言 | 182 |
| 9.2 粒度计算与聚类分析的关系 | 183 |
| 9.3 粒聚类的基本方法 | 186 |
| 9.3.1 模糊聚类分析 | 186 |
| 9.3.2 粗糙集聚类分析 | 188 |
| 9.3.3 商空间聚类分析 | 189 |
| 9.4 基于融合的粒度模型的聚类分析 | 189 |
| 9.4.1 模糊集与粗糙集的结合 | 190 |
| 9.4.2 模糊商空间 | 191 |
| 9.5 多粒度聚类若干问题的研究 | 192 |
| 9.5.1 多粒度聚类中粒子的转换问题 | 192 |
| 9.5.2 约简集粒度的精准性 | 196 |
| 9.5.3 多粒度快速聚类算法 | 197 |
| 9.6 基于多粒度聚类的问题求解应用举例:粗糙 RBF 神经网络的学习算法 .. | 201 |
| 9.6.1 粗糙 RBF 神经网络的学习算法 | 201 |
| 9.6.2 粗糙 RBF 神经网络的可用性与可靠性实验 | 202 |
| 9.7 本章小结 | 205 |
| 参考文献 | 206 |
| 第10章 并行约简与 F-粗糙集 | 210 |
| 10.1 粗糙集基本知识 | 212 |
| 10.2 F -粗糙集 | 214 |
| 10.3 并行约简定义与性质 | 216 |
| 10.4 并行约简算法 | 217 |
| 10.4.1 基于属性重要度矩阵的并行约简算法 | 218 |
| 10.4.2 基于属性重要度矩阵的并行约简算法的优化 | 220 |
| 10.4.3 基于 F -属性重要度的并行约简算法 | 220 |
| 10.4.4 (F, ϵ) -并行约简 | 222 |
| 10.5 决策系统的分解 | 223 |
| 10.6 本章小结 | 226 |
| 参考文献 | 227 |

| | |
|------------------------------------|-----|
| 第 11 章 单调性分类学习 ····· | 229 |
| 11.1 引言····· | 229 |
| 11.2 基于优势关系粗糙集的单调性分类分析····· | 230 |
| 11.3 基于模糊偏好粗糙集的单调性分类分析····· | 232 |
| 11.4 基于排序熵模型的单调性分类分析····· | 239 |
| 11.4.1 Shannon 信息熵 ····· | 240 |
| 11.4.2 有序信息熵 ····· | 240 |
| 11.5 基于排序熵的单调性决策树····· | 244 |
| 11.5.1 程序描述 ····· | 246 |
| 11.5.2 性质研究 ····· | 248 |
| 11.5.3 在人工数据上的实验 ····· | 250 |
| 11.6 本章小结····· | 252 |
| 参考文献····· | 252 |
| 第 12 章 不确定性研究中若干问题的探讨 ····· | 254 |
| 12.1 隶属度的不确定性问题····· | 254 |
| 12.2 运算法则的不确定性问题的····· | 256 |
| 12.3 模糊运算与逻辑运算问题····· | 260 |
| 12.3.1 模糊运算 ····· | 260 |
| 12.3.2 逻辑运算 ····· | 262 |
| 12.4 排序的不确定性问题的····· | 265 |
| 12.5 截集水平的不确定性问题的····· | 268 |
| 12.6 Fuzzy 集合的互补律问题的····· | 268 |
| 12.7 集合的统一问题的····· | 269 |
| 12.8 本章小结····· | 270 |
| 参考文献····· | 270 |
| 第 13 章 基于云模型的文本分类应用 ····· | 273 |
| 13.1 云模型在文本挖掘中的理论扩充····· | 273 |
| 13.1.1 基于 VSM 模型的文本知识表示 ····· | 273 |
| 13.1.2 基于信息表的文本知识表示 ····· | 274 |
| 13.1.3 基于云模型的文本信息表转换 ····· | 275 |
| 13.1.4 基于云相似度的文本相似度量 ····· | 276 |
| 13.2 文本分类及其常用方法····· | 278 |
| 13.2.1 文本分类概述 ····· | 278 |
| 13.2.2 文本分类常用方法 ····· | 279 |

| | | |
|---------------|--------------------------|------------|
| 13.2.3 | 性能分析 | 282 |
| 13.2.4 | 文本分类模型的评估 | 283 |
| 13.3 | 基于云模型与粒计算的文本分类 | 285 |
| 13.3.1 | 虚拟泛概念树及概念跃升 | 285 |
| 13.3.2 | 基于云模型的文本特征自动提取算法 | 287 |
| 13.3.3 | 基于云概念跃升的文本分类 | 297 |
| 13.4 | 本章小结 | 300 |
| | 参考文献 | 301 |
| 第 14 章 | 数据挖掘算法的云实现 | 303 |
| 14.1 | 在云上实现数据挖掘算法的技术背景 | 303 |
| 14.2 | 现有基于云计算的数据挖掘平台 | 304 |
| 14.2.1 | “大云”系统 | 304 |
| 14.2.2 | Mahout 开源项目 | 304 |
| 14.2.3 | 电子科技大学与华为公司合作的云挖掘项目 | 305 |
| 14.3 | 经典数据挖掘算法的 MapReduce 实现思路 | 305 |
| 14.4 | 经典数据挖掘算法在 Hadoop 平台的实现范例 | 307 |
| 14.4.1 | 协同过滤算法在 Hadoop 平台的实现 | 307 |
| 14.4.2 | 朴素贝叶斯算法在 Hadoop 平台的实现 | 312 |
| 14.5 | 云挖掘技术的展望 | 318 |
| 14.5.1 | 针对 Web 信息的云挖掘 | 318 |
| 14.5.2 | 针对图结构的云挖掘 | 319 |
| 14.5.3 | 针对声音与视频等多媒体信息的云挖掘 | 319 |
| | 参考文献 | 319 |

第1章 双向认知计算模型——云模型

王国胤^{1,2,3} 许昌林^{1,2} 张清华¹

1. 重庆邮电大学 计算智能重庆市重点实验室

2. 西南交通大学 信息科学与技术学院

3. 中国科学院重庆绿色智能技术研究院 电子信息技术研究所

人类的认知,实质上是客观世界的一种映像。客观世界的不确定性,决定了人类主观认知过程的不确定性。这种不确定性通过认知的最小(基本)单元——概念反映出来。因此,人类在认知过程中的概念形成和知识更新,不可避免地伴随着不确定性的产生。不确定性概念具有随机性、模糊性、不完备性和不稳定性等特征。在不确定性的研究中涌现出许多理论和方法,其中,研究随机性的概率论至今已有百余年的历史,并形成概率论、随机过程和数理统计3大分支。20世纪中后期,以模糊集合为代表研究模糊性、不完备性等知识表示和推理的理论和方法陆续出现,推动了不确定性研究的发展。粒计算是研究和模拟人类从不同侧面、不同粒度对事物进行表示、分析和推理的方法。模糊集、粗糙集和商空间是目前粒计算的三大主要方法,并应用于机器学习、数据挖掘和智能控制等领域。当前,大多数数据挖掘和机器学习的理论和方法主要是研究从定量数据中提取定性概念,也就是在从概念外延(客观世界中的样本集合)抽取概念内涵(主观世界中的抽象概念)的研究。而要使机器具有或者能够模仿人的智能,人类和机器之间能够相互“理解”,相互“交流”,首先要使人类语言所描述的定性知识(概念)能被计算机所理解和处理,这不可避免地要建立从定性概念到定量数据的转换模型,实现定性概念与定量数据的相互转换,即概念内涵与外延的相互转换。云模型是一个在概率论和模糊数学理论两者结合的基础上研究定性概念与其定量表示的认知计算模型。云模型用期望、熵和超熵这三个数字特征来定量描述一个不确定性概念,通过正向云发生器和逆向云发生器实现定性概念与定量数据之间的双向认知变换,揭示客观对象具有的模糊性和随机性,刻画人类的双向认知计算过程。本章将把云模型作为一种实现概念内涵与外延双向转换这一认知计算的理论模型,讨论其一般递归定义形式,并对正向云发生器和逆向云发生器模拟实现的双向认知计算过程进行分析和讨论。

1.1 引言

认知是一个源自于心理学的概念,《辞海》将认知解释为人类认识客观事物、获得知识的活动,包括知觉、记忆、学习、语言、思维和问题解决等过程。人类对事物的认知,往往是从一个“不知”到“了解”,再到“理解”的过程。人脑接受外界输入的信息,经过头脑的加工处理,转换成内在的心理活动,再进而支配人的行为,这个过程就是信息加工的过程,也就是认知过程^[1]。认知科学是研究人类感知和思维对信息处理过程的科学,包括从感觉的输入到复杂问题的求解,从人类个体到人类社会的智能活动,以及人类智能和机器智能的性质^[2],它是现代心理学、信息科学、神经科学、数学、科学语言学、人类学乃至自然哲学等学科交叉发展的结果。认知科学是20世纪世界科学标志性的新兴研究门类,它作为探究人脑或心智工作机制的前沿性尖端学科,已经引起了全世界科学家们的广泛关注。认知科学的研究将使人类自我了解和自我控制,把人的知识和智能提高到前所未有的高度。生命现象错综复杂,许多问题还没有得到很好的说明,而能从中学习的内容也是大量的、多方面的。如何从中提炼出最重要的、关键性的问题和相应的技术,这是许多科学家长期以来追求的目标^[3]。认知计算源自模拟人脑的计算机系统的人工智能,20世纪90年代后,研究人员开始用认知计算一词,用于教计算机像人脑一样思考,而不只是开发一种人工系统^[4]。传统的计算技术是定量的,并着重于精度和序列等级,而认知计算则试图解决生物系统中不精确、不确定和部分真实的问题。认知计算是认知科学的子领域之一,也是认知科学的核心技术领域。认知计算对于未来信息技术、人工智能等领域均有着十分重要的影响。研究认知的机理,建立认知的模型,然后用计算机模拟人类认知的过程来处理实际问题是人工智能领域的重要课题,受到很多研究者的关注^[4-5]。

哲学上说,运动是物质的根本属性,这些永恒运动着的物质所带来的不确定性存在于我们生活的各个角落。自从人类认识到世界的这一本质规律之后,便开始了漫长的探索,试图更加深入地了解客观存在的不确定性,以科学的语言去掌握其中的规律,从而达到认知世界的目的^[6]。认知的不确定性,顾名思义反映的是人的主观意识在对客观世界的认识过程中所存在的不确定性,这种不确定性是由客观世界本身的不确定性同人类自身认知能力的不确定性共同决定的,是通过认知的最小(基本)单元——概念反映出来的。对同一个事件,由于不同人的认知能力差异,观察角度不同,会表现出对概念表达的差异。认知的不确定性,归根到底来源于客观世界的不确定性,客观世界的不确定性映射到人脑形成的概念(主观世界)也必然是不确定的^[7-8]。因此,人类在认知过程中的概念形成和知识更新,不可避免地伴随着不确定性的产生^[9]。而概念的不确定性必然导致计算和推理的不确定性^[10]。研究不确定性概念的表达、处理和

模拟,寻找不确定性概念的内涵特征和外延特征,并进行形式化表示,用机器模拟研究人类认知不确定性概念的认知过程,成为当前人工智能领域研究的重要任务^[11-14]。

不确定性概念具有随机性、模糊性、不完备性和不稳定性等特征。其中,随机性和模糊性是最基本的特征。随机性是由于条件不能决定结果而表现出来的不确定性,它反映了因果律的缺失问题;模糊性是由于概念外延边界的不清晰而表现出来的不确定性,它反映了排中律的缺失问题,也就是由于概念外延的不分明性、事物对概念归属的亦此亦彼性等。Zadeh 教授于 1965 年创建的模糊集合论^[15]、Pawlak 教授于 1982 年建立的粗糙集理论^[16]、Atanassov 教授于 1986 年提出的直觉模糊集^[17]以及近年来 Mendel 教授提出的二型模糊集^[18]等都是处理模糊性问题的有力工具。模糊集理论通过隶属函数刻画不确定性概念的模糊性,粗糙集通过上、下近似集合来处理模糊性,直觉模糊集通过对模糊对象赋予真、假隶属函数来处理模糊性,二型模糊集研究隶属度本身的模糊性问题。然而,模糊集可以由不同阈值的截集来确定,模糊集的截集却是普通的康托集,是确定的集合。因此,模糊集与它的截集体现了模糊概念和清晰概念在不同粒度层次上相互转换的辩证关系。从粒计算的观点来看,概念具有的不确定性和确定性是概念外延在不同知识粒度层次上的不同表现形式,它们并非完全对立,在一定粒度层次上是可以相互转化的^[19]。

概率论、模糊集理论、粗糙集理论从不同的角度研究了概念的不确定性问题。概率论重在研究概念外延的随机性特征;模糊集理论用隶属函数来刻画概念外延亦此亦彼的程度;粗糙集理论通过上、下近似集来描述不确定性概念外延的边界域,是一种用精确的方法解决不确定性概念外延边界线的方法。这些研究方法虽然有着明确的切入点 and 严谨的约束条件,但也存在着一定的局限性。比如,概率论的一个基本假设就是基本事件的概率之和严格等于 1,但对于自然语言中使用的概念而言,未必严格如此。另外,对于随机变量,它首先需要确定其分布类型和分布参数,这就需要大量的样本,目前一般都是事先假定其服从某种类型的分布,并估计分布参数,因此,应用随机变量来研究不确定性概念,有时会产生较大的误差。在模糊集理论中,对于模糊变量,关键在于确定其隶属函数,而隶属函数的确定往往带有主观性。在粗糙集理论中,对于粗糙变量,最大的优势是它的客观性,上近似集、下近似集、属性重要度、信息熵等都可以直接从数据中计算获得,不需要人为主观参与,但它处理的数据样本无法体现整体数据的分布,丢失了数据样本的随机特性,其泛化能力有待提高,而且粗糙集在知识约简过程中存在的过度学习问题也需要解决^[20-21]。

当前,人们的研究重点集中在从概念外延(客观世界中的样本集合)到概念内涵(主观世界中的抽象概念)变换的研究,目前众多机器学习和数据挖掘研究者所开展的研究工作,已经取得很多研究成果,推动了人工智能的发展;但从概念内涵(抽象概念)到概念外延(样本集合)的研究非常困难,研究成果甚少,其关键原因在于由于认知个

体的差异,特别是概念内涵本身的不确定性(主要是模糊性),导致了概念外延的不确定性。不同的人针对同一抽象概念(模糊概念)可能会得到不同的外延(客观对象集)。因此,研究如何形式化地表示不确定性概念、用概念外延描述概念内涵、从概念外延挖掘得到其内涵等问题成为认知计算领域需要研究的关键问题。

在从不确定性概念内涵到外延映射的相关研究中,谓词逻辑创始人 Frege 教授早在 1904 年就提出了“含糊”一词,并把它归结为概念外延边界域问题。模糊集的提出让不少逻辑学家试图用这一理论解决 Frege 教授的含糊概念问题,但遗憾的是模糊集无法计算出它的边界线上具体的含糊对象的数目。Pawlak 教授针对 Frege 教授的边界域思想,把那些无法确认的个体都归属于边界域,而这种边界域被定义为上近似集与下近似集的差集。由于上近似集与下近似集都可以通过等价关系给出确定的数学描述公式,所以含糊元素数目可以被计算出来,从而实现了 Frege 教授的边界域思想。但是,粗糙集理论没有刻画对象本身具有的随机性。1995 年,李德毅教授在概率论和模糊数学理论两者结合的基础上提出的云模型理论融合了数据分布的特点,利用期望、熵和超熵这三个数字特征来定量描述一个不确定性概念,通过正向云发生器实现从概念的内涵向其外延进行转化,通过逆向云发生器实现从概念的外延向其内涵进行转化,较好地刻画了概念内涵与外延之间的双向认知变换过程,同时也揭示了客观对象具有的模糊性和随机性^[11-14,22]。

在云模型论中,基于正态分布和正态隶属函数的正态云是一种重要的云模型,具有普适性^[12]。本章将以正态云模型为代表研究人类认知的双向认知计算过程,并对同一个人对同一个定性概念以及不同人对同一个定性概念的双向认知计算过程进行实验模拟和分析。

1.2 正态分布与正态隶属函数

正态分布广泛存在于自然现象、社会现象、科学技术和生产活动中,现实世界中的许多随机现象都服从或者近似服从正态分布。例如,正常生产条件下的产品质量指标,随机测量误差,同一生物群体的某种特征,某地的年平均气温等。正态分布是许多重要概率分布的极限分布,许多非正态的随机变量是正态随机变量的函数,正态分布的密度函数和分布函数有各种很好的性质和比较简单的数学形式,这些都使得正态分布在理论和实际中应用非常广泛^[23]。

中心极限定理从理论上阐述了产生正态分布的条件^[23]。对中心极限定理的简单直观说明是:如果决定某一随机变量结果的是大量微小的、独立的随机因素之和,并且每一因素的单独作用相对均匀地小,没有一种因素可起到压倒一切的主导作用,那么这个随机变量一般近似于正态分布。例如,某种成批生产的产品,如果工艺、设