

不确定数据信任分类 与融合

刘准钊 潘 泉 [法] Jean Dezert 周 旷 焦连猛 著

 科学出版社

不确定数据信任分类与融合

刘准钊 潘 泉 [法] Jean Dezert 周 旷 焦连猛 著

科学出版社

北 京

内 容 简 介

本书面向高层(决策层)信息融合,深入系统地介绍作者在不确定数据信任分类与多源信息融合推理方面的重要研究进展。首先,介绍一种全新的基于证据推理框架的信任分类模型,其主要处理不确定信息建模分析问题。面向不同情况,本书讲述如不确定数据信任 C 均值/中值聚类算法、关系数据多中心信任聚类算法、信任 K 近邻分类器、快速信任分类器以及不完整数据信任分类器等多种方法,将多个信任分类器通过融合推理能够得到更全面可靠的信息,提高判断决策的准确性。因此,本书还介绍信源可靠度不同情况下的自适应加权折扣融合方法,解决高冲突信息融合问题。并且,介绍面向多源多时相信息融合发展动态证据推理,解决目标识别与变化检测一体化问题。

本书内容力求系统性和实用性,覆盖作者在该领域最新的研究进展,可以为从事信息融合、机器学习及模式识别等领域的科技工作者提供参考借鉴。

图书在版编目(CIP)数据

不确定数据信任分类与融合/刘淮钊等著. —北京:科学出版社,2016.3

ISBN 978-7-03-047501-5

I.①不… II.①刘… III.① 人工智能—模式识别—研究 IV.①TP18

中国版本图书馆 CIP 数据核字(2016) 第 044155 号

责任编辑:李 萍 纪四稳/责任校对:胡小洁

责任印制:赵 博/封面设计:红叶图文

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>

天津新科印刷厂 印刷

科学出版社发行 各地新华书店经销

*

2016 年 3 月第 一 版 开本:720×1000 1/16

2016 年 3 月第一次印刷 印张:16 1/4

字数:328 000

定价:95.00 元

(如有印装质量问题,我社负责调换)

前 言

在错综复杂的战场环境下,由于外界干扰、人为对抗欺骗,以及传感器自身性能等诸多因素的影响,信源信息一般具有很大的随机性、不精确性、不完备性以及高冲突性等特点。在处理不确定信息时,单一传感器难以保证对不同类型的目标都具有较好的探测识别能力,并且单一传感器对复杂的自然环境和电磁环境缺乏较强的适应性,因而信息融合技术在战略预警等复杂军事系统中有着广泛的应用。通过对不同传感器的多源信息综合处理能够得到更全面可靠的信息,这样可以提高目标探测的准确性,增强系统的鲁棒性。在高层(决策层)信息融合中,一般是各传感器先通过对自身观测数据的预处理和特征提取等得到对目标的初步分类判断,然后在融合中心对各传感器的分类信息进行融合以获得准确可靠的目标识别结果或变化检测结果。作为信息融合一种重要的基本理论方法,证据推理将概率中由单元素构成的辨识框架拓展到包含单元素及其集合的幂集,并定义了一系列信任函数和融合规则,这使其在不确定信息的分析处理上具有独特优势。研究基于证据推理的高层信息融合包括不确定数据信任分类和多源信息融合推理方法具有重要的理论意义和应用价值。

作者长期从事证据推理及其应用的研究工作,关于证据推理在数据分类和融合推理方面的最新研究进展多散布于学术文献,目前国内外鲜有系统、全面介绍不确定数据信任分类和融合推理的著作。对于广大对不确定信息融合推理有兴趣的学者,拥有一部反映证据推理及其在不确定数据分类与融合方面研究进展的工具书,将使他们的努力取得事半功倍的效果;对于长期在工程第一线从事具体工作的工程技术人员,往往也需要一本反映研究发展动态的参考书,以便为更新知识而充电。本书的初衷正是试图为他们尽微薄之力,希望对他们有所帮助。

作者是在多年证据推理及其应用研究工作基础上完成本书的。本书较为系统地总结了国内外在本领域的最新研究进展,特别是作者在不确定数据聚类、分类,不完整数据分类,多源信息加权融合和动态证据推理等方面的研究成果。本书共12章,包括不确定数据信任分类和多源信息融合推理两大部分内容。第1章绪论主要介绍信息融合、证据推理和不确定数据分类、聚类等基础知识,对相关研究工作进行总结;第2章介绍不确定数据信任 C 均值聚类算法,在没有训练样本的情况下将传统模糊 C 均值聚类算法拓展到证据推理框架下;第3章介绍信任 C 中值聚类算法,主要讲述关系数据集在证据推理框架下的信任划分;第4章介绍面向复杂关系数据集的多中心信任聚类算法,主要研究由多个样本点加权构建类中

心的聚类方法;第5章介绍不确定数据广义信任 K 近邻分类器,主要论述基于 K 近邻的基本信任指派构造和多源证据自动融合;第6章介绍一种较为简便的基于类中心的不确定数据广义信任分类器,主要研究基于类中心的基本信任指派构造和融合;第7章介绍一种适用于大数据量的快速广义信任分类器,主要论述利用类特征向量来直接计算目标数据的广义信任分类;第8章介绍一种鲁棒的信任规则分类系统,主要研究利用信任规则对不确定数据分类问题进行建模和推理;第9章介绍同时考虑不确定训练数据与专家知识的复合信任规则分类系统,可以实现对异类不确定信息的综合利用;第10章介绍带有属性信息缺失的不完整数据广义信任分类方法,包括缺失值的估计以及多分类结果的融合等;第11章介绍多源信息加权融合算法,主要研究证据间不一致性的度量、证据加权折扣融合等,并应用到多源信息融合目标识别;第12章介绍动态证据推理,定义新的推理空间、信任函数等,设计动态证据融合规则,并将其应用到多源遥感图像变化检测。相信本书的出版对于推动证据推理及应用的研究发展会有所裨益。

本书工作得到了国家自然科学基金重点项目“基于联合决策与估计的高频超视距雷达信息处理与融合”(61135001)、国家自然科学基金青年项目“带有属性信息缺失的不完整数据广义信任分类”(61403310),以及陕西省自然科学基金(2015JQ6265)和中央高校基本科研业务费专项资金(3102014JCQ01067)等项目资助,在此,对以上单位表示衷心的感谢。感谢法国 Telecom Bretagne 的 Gregoire Mercier 教授!Gregoire 是本书作者刘准钊在法国的博士生导师,他对本书中动态证据推理部分的研究工作给予了认真的指导,提出了很多建设性意见。感谢法国 University of Rennes 1 的 Arnaud Martin 教授在信任中值聚类方面给予了重要的指导。感谢法国 University of Technology of Compiègen 的 Thierry Denoeux 教授在信任规则分类方面给予的重要指导。感谢课题组的程咏梅教授,程教授把本书作者刘准钊引入证据推理及应用研究的道路,她的悉心指导和严格锤炼对作者终生受益。感谢海军航空工程学院的何友院士、南京电子技术研究所的贾德院士和北京理工大学的毛二可院士,他们审阅了书稿,提出了宝贵的建议,并给本书撰写了推荐意见。感谢课题组的张洪才教授、梁彦教授、张绍武教授、赵永强教授、赵春晖副教授、杨峰副教授、王小旭副教授、刘勇博士、王增福博士、徐林峰博士、兰华博士、李琳博士等,本书的许多研究工作曾得到他们的启发和帮助。感谢和我们朝夕相处、共同研究、共同成长的研究生,吕洋、康青青、王莉、李立、祝志勇、孙帅、周平、刘永超等研究生参与了本书部分章节的整理工作,正是他们的热心帮助才使本书能够顺利出版。由于作者水平有限,书中不妥之处在所难免,恳请读者批评指正。

作者

2015年11月

目 录

前言

第 1 章 绪论	1
1.1 战略预警中的信息融合	1
1.2 证据推理简介	3
1.3 不确定数据分类	5
1.3.1 不确定数据无监督聚类	5
1.3.2 不确定数据有监督分类	10
1.3.3 不完整数据分类	14
1.4 多源不确定数据融合	16
1.4.1 证据融合规则	17
1.4.2 证据修正法	20
第 2 章 不确定数据信任 C 均值聚类算法	23
2.1 引言	23
2.2 证据 C 均值算法简述及分析	25
2.3 信任 C 均值聚类算法	27
2.3.1 BCM 的基本原理	27
2.3.2 BCM 的目标函数	28
2.3.3 最小化目标函数 J_{BCM}	30
2.4 关系信任 C 均值聚类算法	33
2.5 信任划分的近似	35
2.6 实验分析	36
2.6.1 珍珠数据	36
2.6.2 人工合成数据示例	37
2.6.3 真实数据	42
2.6.4 遥感图像分类	43
2.7 本章小结	45
第 3 章 关系数据信任 C 中值聚类算法	46
3.1 引言	46
3.2 信任 C 中值聚类	46

3.2.1	目标函数	46
3.2.2	目标函数优化	48
3.2.3	模型参数	50
3.3	基于 MECM 的网络社区发现	51
3.3.1	信任模块函数	51
3.3.2	基于 MECM 的社区检测算法	52
3.4	实验分析	52
3.4.1	重叠数据集	55
3.4.2	高斯数据集	59
3.4.3	人工合成网络数据集	62
3.4.4	真实网络数据集	64
3.5	本章小结	67
第 4 章	关系数据多中心信任聚类算法	69
4.1	引言	69
4.2	多中心模糊聚类简介	70
4.3	多中心信任聚类算法	71
4.3.1	基于单中心的信任聚类	71
4.3.2	基于样本加权的多中心信任聚类	73
4.4	实验分析	76
4.4.1	重叠数据集	76
4.4.2	X_{12} 数据集	79
4.4.3	X_{11} 数据集	80
4.4.4	空手道俱乐部	82
4.4.5	UCI 数据集	84
4.5	本章小结	86
第 5 章	不确定数据广义信任 K 近邻分类器	88
5.1	引言	88
5.2	证据 K 近邻算法简介及分析	90
5.3	广义信任 K 近邻分类器	91
5.3.1	构造基本信任指派	91
5.3.2	融合基本信任指派	94
5.3.3	参数的优化	97
5.3.4	BK-NN 的分类表达能力分析	98
5.4	实验分析	99

5.4.1	人工合成数据示例	99
5.4.2	真实数据	103
5.4.3	目标识别数字算例	105
5.5	本章小结	106
第 6 章	基于类中心的不确定数据广义信任分类器	108
6.1	引言	108
6.2	基于类中心的广义信任分类方法	109
6.2.1	子分类结果计算	109
6.2.2	多分类结果融合	110
6.2.3	参数调整规则	114
6.3	实验分析	116
6.3.1	人工合成数据示例	117
6.3.2	仿真分析	119
6.3.3	真实数据	120
6.4	本章小结	121
第 7 章	不确定数据快速广义信任分类方法	122
7.1	引言	122
7.2	快速广义信任分类规则	122
7.2.1	估计各类中心向量	122
7.2.2	计算基本信任指派	126
7.3	仿真实验分析	128
7.3.1	实验 1: 两类问题仿真数据	128
7.3.2	实验 2: 四类问题仿真数据	129
7.3.3	实验 3: 大数据量样本仿真实验	130
7.3.4	实验 4: 真实数据集	131
7.4	本章小结	132
第 8 章	不确定数据信任规则分类方法	134
8.1	引言	134
8.2	信任规则分类系统	135
8.2.1	信任规则结构	136
8.2.2	基于数据驱动的信任规则库构建	137
8.2.3	信任推理方法	140
8.3	实验分析	142
8.3.1	实验设置	142

8.3.2	分类正确率评估	144
8.3.3	分类鲁棒性评估	147
8.3.4	时间复杂度分析	151
8.4	本章小结	155
第 9 章	不确定训练数据与专家知识复合信任规则分类方法	156
9.1	引言	156
9.2	复合信任规则分类系统	157
9.2.1	基于知识驱动的信任规则库构建	157
9.2.2	复合信任规则库构建	160
9.3	基于不确定训练数据与专家知识的多源目标识别	163
9.3.1	问题描述	163
9.3.2	目标识别复合信任规则库构建	163
9.3.3	对比分析	167
9.3.4	参数分析	169
9.4	本章小结	170
第 10 章	不完整数据广义信任分类器	171
10.1	引言	171
10.2	不完整数据广义信任分类算法	172
10.2.1	缺失属性值的估计	173
10.2.2	带有估计值的样本分类	174
10.2.3	c 个折扣分类结果的融合	175
10.3	仿真实验	180
10.3.1	人工合成数据示例	180
10.3.2	仿真分析	183
10.3.3	真实数据	184
10.4	本章小结	185
第 11 章	多源信息自适应折扣融合	186
11.1	引言	186
11.2	证据距离	187
11.2.1	Jousselman 距离和 Bhattacharyya 距离	187
11.2.2	基于概率框架的证据距离	188
11.3	证据间的内在冲突	191
11.4	新的证据不一致性量测	193
11.5	多源证据折扣因子的确定	195

11.6	多传感器目标识别数字算例	197
11.7	基于 ISR 仿真实验平台的多源信息目标识别	201
11.7.1	基于 HLA 的 ISR 仿真实验平台构架及联邦设计	201
11.7.2	多源决策层信息融合模型	204
11.7.3	多源信息融合目标识别仿真	204
11.8	本章小结	209
第 12 章	动态证据推理	211
12.1	引言	211
12.2	动态证据推理	212
12.2.1	状态转移空间	212
12.2.2	动态证据推理融合规则	215
12.2.3	先验信息下的条件推理	219
12.2.4	动态证据推理的决策支持	220
12.3	变化检测数字算例	221
12.4	基于动态证据推理的遥感图像变化检测	226
12.4.1	简化的信任 C 均值聚类算法	227
12.4.2	两幅异质遥感图像 (ERS & SPOT) 融合	230
12.4.3	三幅序列遥感图像的融合	233
12.5	本章小结	237
参考文献		238

第1章 绪 论

在错综复杂的战场环境(如战略反导预警系统)下,单一传感器难以保证对不同类型的目标都具有较好的探测识别能力,并且单一传感器对复杂的自然环境和电磁环境缺乏较强的适应性,因而信息融合技术在战略预警系统中有着广泛的应用。通过对不同传感器的多源信息综合处理能够得到更全面可靠的信息,这样可以提高目标探测的准确性,增强系统的鲁棒性。反导预警系统中包括天基、地基、空基、海基等多种传感器平台,传感器属性及工作原理都不相同,且不同平台间信息需要传输,因而一般采取适用于异质传感器且对通信带宽要求较低的决策层信息融合。即各传感器先通过对自身观测信息数据的预处理和特征提取等得到对目标的初步判断,然后在融合中心对各传感器的分类判断信息进行融合以获得准确可靠的目标识别结果或变化检测结果。

由于外界干扰、人为对抗欺骗,以及传感器自身性能等诸多因素的影响,信源(传感器)信息一般具有很大的不确定性和高冲突性^[1, 2]。这里的不确定性(uncertainty)是指信息在噪声等影响下所具有的随机性(randomness)和属性知识不充分,以及缺失情况下表现出的不精确性(imprecise)。传统的概率框架主要表示信息的随机性,而在对不精确信息建模分析时存在较大局限^[3-7]。证据推理^[8-11]一般被视为概率理论的拓展,其将概率论中由单元素构成的辨识框架 Ω 拓展到包含单元素及其集合(如辨识框架的任意子集)的幂集 2^Ω ,并在幂集下定义了一系列信任函数及融合规则。证据推理在处理不确定信息(包含随机性和不精确性的信息)时具有独特的优势。

1.1 战略预警中的信息融合

战略预警是一个国家安全和战略军事行动的重要保障,是国家战略防御和威慑力量不可缺少的重要基础。战略预警对象主要包括弹道导弹、战略轰炸机、远程巡航导弹、高空长航时无人机、空天飞机、平流层浮空器、空间目标等。战略预警系统主要由预警指挥控制中心和空间预警卫星网、地面预警雷达网、预警飞机及预警通信网等构成,一般具有如下三大特点。

(1) 平台多样,包括天基、地基、空基、海基等探测系统,武器系统,指控系统等。

(2) 系统集成,各系统在通信网络的支持下,在信息空间的各个层次、时间各

宽(传感器处于不同平台,需数据传输)和计算机资源要求较高,且受环境等因素的制约较大。

特征层融合^[12-14,16]是指传感器利用各自获取的观测数据进行特征提取,融合中心对各传感器提供的特征向量进行融合并根据融合特征对目标身份做出识别判断。特征融合前需要对各传感器的特征进行关联处理,保证参与融合的特征矢量来自同一个目标,这可以通过目标状态估计技术来解决。特征层融合对数据配准的要求比数据层融合稍微宽松,参与融合的传感器可以是异质的,具有较大的灵活性。但是,各传感器提供的特征数据量纲可能不统一,且融合特征的维数一般较高,这给后续的模式识别带来了困难。

决策层融合^[12-14,16]是指各传感器先在本地对观测信息分别进行预处理、特征提取、模式分类,得到各自对目标身份的初步识别,然后将各传感器的识别结果进行关联,确保参与融合的识别结果来自同一个目标,这可以通过目标状态估计技术来解决。融合中心对关联后的各传感器识别结果进行融合以获得最终的识别结果。决策层融合中虽然原始信息的损失较多,但是在信息处理方面具有很高的灵活性,对通信带宽要求较低,并且对传感器的依赖性小,传感器属性既可以同质,也可以异质,还能处理异步信息。

在战略预警系统中,处于不同平台的传感器(如空间预警卫星、地面预警雷达等)之间在数据传输时对通信要求较高,且各传感器的属性可能同质,也很可能异质。因此,决策层融合在战略预警系统中有着广泛的应用,也是战略预警系统中的一个研究热点。本书将研究基于决策层融合的多源信息目标识别和目标变化检测。在目标识别方面,主要是想利用多源信息之间的互补性和冗余性来提高目标识别准确度,增强目标识别系统对外界环境和干扰的鲁棒性。目标变化检测主要通过对多源多时相信息的融合来检测出信息的差异性,这可应用于战场毁伤评估等方面。

在错综复杂的战场环境中,信源信息具有很大的不确定性和高冲突性。目前,在多源信息融合中有大量的不确定信息处理方法^[1,5,13-15],其主要基于贝叶斯概率论、粗集、模糊集、可能性理论、灰色系统理论和证据推理。其中,证据推理将概率框架拓展到了幂集,并发展了众多的证据融合方法,这使证据推理在对不确定信息建模处理上具有独特的优势。

1.2 证据推理简介

证据推理(evidence reasoning)是由 Dempster 在 1967 年最先提出的,他用多值映射得出了概率的上下界^[17],后来由 Shafer 在 1976 年推广并且形成证据推理^[9],因此又称 Dempster-Shafer 理论(Dempster-Shafer theory, DST)。

在 Shafer 模型下,定义一个包含有限个互斥且完备的元素集合 $\Omega = \{\omega_1, \dots,$

$\omega_n\}$ 为所研究问题的辨识框架 (frame of discernment)。由于 Shafer 模型对辨识框架 Ω 所包含元素的互斥性和完备性的要求, 所以 Shafer 模型是一个假设的封闭模型 (closed-world assumption)。

Ω 的所有子集构成的集合称为 Ω 的幂集, 表示为 2^Ω , 其包含了 $2^{|\Omega|}$ 个元素。这里 $|A|$ 表示集合 A 的势, 即其所包含的元素个数 (如 $|\{\omega_1, \omega_2\}| = |\omega_1 \cup \omega_2| = 2$)。例如, 如果 $\Omega = \{\omega_1, \omega_2, \omega_3\}$, 那么 $2^\Omega = \{\emptyset, \omega_1, \omega_2, \omega_3, \omega_1 \cup \omega_2, \omega_1 \cup \omega_3, \omega_2 \cup \omega_3, \Omega\}$ 。

在 Shafer 模型下, 一个证据的基本信任指派 (basic belief assignment, BBA) 就是从 2^Ω 到 $[0, 1]$ 上的一个映射函数 $m(\cdot) : 2^\Omega \rightarrow [0, 1]$ (又称基本信任函数或 mass 函数), 并满足以下条件

$$\begin{cases} \sum_{A \in 2^\Omega} m(A) = 1 \\ m(\emptyset) = 0 \end{cases} \quad (1.1)$$

如果 $m(A) > 0$, 则 A 称为焦元 (focal element)。如果 $m(A) = \max(m(\cdot))$, 则称 A 为主焦元。一个基本信任指派 $m(\cdot)$ 中所有焦元的集合构成了 $m(\cdot)$ 的核 (core/kernel), 表示为 $k(m)$ 。 $m(A)$ 表示的是目标属于命题 A 的基本信任值, 而不包含 A 的子集。

在辨识框架 Ω 下, 一个目标样本 x 允许以不同的基本信任值属于幂集 2^Ω 中的各个元素, 其包括单类 (如 ω_i) 和若干单类的集合 ($\omega_i \cup \dots \cup \omega_j$, 又称复合类), 其基本信任指派函数为 $m(\cdot)$, 这种分类方法称为广义信任分类。在概率框架下的分类方法, 仅允许目标以不同的概率属于辨识框架 Ω 中的各个单类, 而未考虑复合类。近来由 Denoeux 发展的基于证据推理的分类方法仅考虑了用所有单类的集合即整个辨识框架来表示未知信息, 未考虑其他个别单类的集合即局部未知信息, 所以这种方法称为狭义信任分类。

Shafer 还在基本信任指派的基础上定义了信任函数 $\text{Bel}(\cdot)$ 和似然函数 $\text{Pl}(\cdot)$, 即

$$\text{Bel}(A) = \sum_{A, B \in 2^\Omega; B \subseteq A} m(B) \quad (1.2)$$

$$\text{Pl}(A) = \sum_{A, B \in 2^\Omega; A \cap B \neq \emptyset} m(B) = 1 - \text{Bel}(\bar{A}) \quad (1.3)$$

式中, $\bar{A}$ 表示元素 A 在集合 Ω 中的补集。

信任函数 $\text{Bel}(\cdot)$ 和似然函数 $\text{Pl}(\cdot)$ 通常用来表示概率上下界^[17], $[\text{Bel}(A), \text{Pl}(A)]$ 可以被解释为 A 的不确定信任区间。 $m(\cdot)$ 、 $\text{Bel}(\cdot)$ 、 $\text{Pl}(\cdot)$ 三者之间是一一对应的关系, 可以相互转换, 有固定的转换关系式^[9]。

来源于两个独立证据且在一个辨识框架下的基本信任函数 $m_1(\cdot)$ 和 $m_2(\cdot)$ 可以利用 Dempster-Shafer 规则 (简称 DS 规则) 进行融合, 以获得融合的基本信任函

数 $m(\cdot) = [m_1 \oplus m_2](\cdot)$ 。而 $A \neq \emptyset, B, C \in 2^\Omega$, 有

$$\begin{cases} m(\emptyset) = 0 \\ m_{\text{DS}}(A) = \frac{\sum_{\substack{B \cap C = A \\ B \cap C = \emptyset}} m_1(B)m_2(C)}{1 - \sum_{B \cap C = \emptyset} m_1(B)m_2(C)} = \frac{\sum_{B \cap C = A} m_1(B)m_2(C)}{\sum_{B \cap C \neq \emptyset} m_1(B)m_2(C)} \end{cases} \quad (1.4)$$

其中两个证据的冲突信息一般定义为

$$k_{12} = \sum_{B \cap C = \emptyset} m_1(B)m_2(C) \quad (1.5)$$

DS 规则将所有冲突信息 k_{12} 都按比例分配给了其他焦元。只有当组合规则式 (1.4) 分母不为零时, 即两个证据不完全冲突 (如 $k_{12} \neq 1$) 时, DS 规则才有效。DS 规则的融合结果 $m(\cdot)$ 的核相当于 $m_1(\cdot)$ 和 $m_2(\cdot)$ 核的交集, 即 $k(m) = k(m_1) \cap k(m_2)$ 。DS 规则的主要优势在于计算简便, 并且满足结合律和交换律, 因此 DS 规则很容易被推广到对 N 个独立且等权重证据的融合。

1.3 不确定数据分类

利用单信源信息对目标进行初步识别实际上是一个不确定数据的模式分类问题。在军事应用中, 由于人为对抗、战场环境复杂多变以及传感器性能限制等, 观测信息还会出现目标部分属性值缺失的情况^[18]。传统不确定、不完整数据分类方法一般基于概率框架。然而, 概率框架仅仅刻画了数据的随机性等特点, 而不能有效地描述数据因属性信息不充分或缺失造成的不精确性。作为概率论的推广, 证据推理通过将辨识框架拓展到幂集, 并定义了概率上下界即信任函数 $\text{Bel}(\cdot)$ 和似然函数 $\text{Pl}(\cdot)$ 等概念, 这样不仅可以很好地表示信息的随机性, 而且也能有效合理地刻画信息的不精确性等特点。目前, 已经有一些简单的基于证据推理的不确定 (完整) 数据分类方法^[19-35], 数据分类根据训练样本 (类别先验) 信息的有无可以分为无监督聚类和有监督分类。同时, 关于有缺失值的不完整数据分类问题, 目前也已经有很多方法^[36]。

1.3.1 不确定数据无监督聚类

数据聚类主要是为了将一个数据集中的 n 个样本 (目标) 在没有任何类别先验信息和训练样本的情况下分为 c 类 $\omega_1, \dots, \omega_c$, 且要保证每一类中的目标在一定的标准下是相似的。目标可以用对象数据 (object data) 或关系数据 (relational data) 来描述^[20, 37]。对象数据利用目标 p 维的自身属性数据来准确表示该目标。关系数据是指目标对之间的相似性或非相似性等特征 (如距离) 量测。目前, 已经有大量关于对象数据和关系数据的聚类方法。数据聚类主要分为分级聚类、硬聚类^[38] 和

模糊聚类^[39, 40]。分级聚类是指利用一系列嵌套的类来描述数据；硬聚类所得到的数据类别具有排他性，是指一个目标一旦属于某个类别就不能再属于其他类别；模糊聚类是指目标以不同的隶属度属于多个类别。其中，最著名的模糊聚类方法就是 Bezdek 的面向对象数据的模糊 C 均值 (fuzzy C -means, FCM) 聚类方法^[39]，以及面向关系数据的关系模糊 C 均值 (relational fuzzy C -means, RFCM) 聚类方法^[41]。

这些模糊聚类方法一般是先设计一个合适的目标函数，然后通过优化 (最小化) 这个目标函数来得到描述所有目标属于各类别的隶属度矩阵 $U_{n \times c} = (u_{ik})$ ，且要满足

$$\begin{cases} \sum_{k=1}^c u_{ik} = 1, i = 1, \dots, n \\ \sum_{i=1}^n u_{ik} > 0, k = 1, \dots, c \end{cases} \quad (1.6)$$

式中， $u_{ik} \in [0, 1]$ 表示目标 i 属于类别 k 的隶属度。有些学者发现模糊聚类对于噪声和野值的鲁棒性较差，往往会得出不合理的结果。因此，提出了放宽对目标隶属度归一化的限制，如式 (1.6)，这也就导致了可能性聚类 (possibilistic partition)。例如，Krishnapuram 等^[40, 42]、武小红等^[43] 学者通过调整目标函数设计了可能性 C 均值 (possibilistic C -means, PCM) 聚类算法。由 PCM 聚类算法得到的目标隶属度 u_{ik} 被解释为目标 i 属于类别 k 的可能性。Dave 等提出了噪声聚类^[44-46]，利用增加的噪声类表示那些很难被传统类别描述的数据点。这些方法最初被提出来适用于处理对象数据，后来也都被推广到处理关系数据。

近些年来，Denoeux 和 Masson 在证据推理框架下提出了信任划分 (credal partition) 的概念^[47, 48]。信任划分是硬划分、模糊划分和可能性划分的一种推广，其将目标分类的框架扩展到了幂集，允许目标以不同的基本信任值不仅可以属于传统的单类，也可以属于单类的集合。信任划分可以更深刻地揭示数据间的关系，且可以提高传统聚类方法对噪声的鲁棒性。在信任聚类算法中，Denoeux 等提出了面向关系数据的 EVCLUS (Evidential CLUSTERing)^[49] 聚类算法，后来在 FCM 和 Dave 的噪声聚类算法的基础上，提出了面向对象数据的证据 C 均值 (evidential C -means, ECM) 算法^[20]。最近，推导了关系证据 C 均值 (relational evidential C -means, RECM) 算法^[37] 来对关系数据进行聚类，同时还将 RECM 算法和 EVCLUS 算法进行了对比，表明 RECM 算法和 EVCLUS 算法的聚类效果相仿，而 EVCLUS 算法的计算量较大。Antoine 等提出了限制证据 C 均值 (constrained evidential C -means, CECM) 聚类算法^[50]，考虑在 ECM 算法中加入一些已知的限制条件，以获得更准确的聚类结果。下面对一些常用的关于 C 均值的聚类算法进行简单介绍。

1. C 均值聚类算法

考虑一个数据集 $\{x_1, \dots, x_n\}$ 包含 n 个目标，其中目标 $x_i (i = 1, \dots, n)$ 是一

个 p 维的向量即 $\mathbf{x}_i \in \mathbb{R}^p$ 。这个数据集来源于 $c(2 \leq c < n)$ 个类别。每个类别对应于一个聚类中心 $\mathbf{v}_k \in \mathbb{R}^p$ ，用 $\mathbf{V}_{c \times p}$ 来表示 c 个聚类中心的向量构成的聚类中心矩阵。

FCM 的目标函数为

$$J_{\text{FCM}}(\mathbf{U}, \mathbf{V}) = \sum_{i=1}^n \sum_{j=1}^c u_{ij}^{\beta} d_{ij}^2 \quad (1.7)$$

在限制条件 (1.6) 下，通过最小化 $J_{\text{FCM}}(\mathbf{U}, \mathbf{V})$ 可以求得目标属于各类的隶属度矩阵 $\mathbf{U}_{n \times c} = (u_{ik})$ 和聚类中心的矩阵 $\mathbf{V}_{c \times p}$ 。这里， β 是隶属度的指数系数，一般取 $\beta = 2$ 。 d_{ij} 指的是目标 $\mathbf{x}_i$ 和聚类中心 $\mathbf{v}_j$ 之间的欧氏距离，即 $d_{ij} \triangleq \|\mathbf{x}_i - \mathbf{v}_j\|$ 。利用迭代方法最小化目标函数，可以分别求得聚类中心 $\mathbf{V}_{c \times p}$ 和隶属度矩阵 $\mathbf{U}_{n \times c} = (u_{ik})$ 。

$$\mathbf{v}_k = \frac{\sum_{i=1}^n u_{ik}^{\beta} \mathbf{x}_i}{\sum_{i=1}^n u_{ik}^{\beta}}, \forall k = 1, \dots, c \quad (1.8)$$

$$u_{ij} = \frac{d_{ij}^{-2/(\beta-1)}}{\sum_{k=1}^c d_{ik}^{-2/(\beta-1)}}, \forall i = 1, \dots, n; \forall j = 1, \dots, c \quad (1.9)$$

FCM 在初始化时，可以随机生成隶属度矩阵或者聚类中心矩阵，然后利用式 (1.9) 和式 (1.8) 迭代更新直到收敛。FCM 算法具有良好的收敛性但可能会导致局部最小^[39]，且 FCM 算法由于限制条件 (1.6) 使其无法检测噪声点和野值点^[40, 42]。实际上，FCM 中目标 i 属于类别 j 的隶属度不仅取决于 d_{ij} ，还和目标到其他所有类别聚类中心的距离有关。Krishnapuram 和 Keller 提出的可能性 C 均值聚类算法^[42]中，放弃了概率的条件限制 (1.6)，且在目标函数中添加了一个惩罚项。

$$J_{\text{PCM}}(\mathbf{U}, \mathbf{V}) = \sum_{i=1}^n \sum_{j=1}^c u_{ij}^{\beta} d_{ij}^2 + \sum_{j=1}^c \eta_j \sum_{i=1}^n (1 - u_{ij})^{\beta} \quad (1.10)$$

式中， $\eta_j > 0$ 是个固定值，由用户来确定，用来平衡目标函数中的两部分。利用和 FCM 相类似的优化过程可以推导出 PCM 算法中用于更新的隶属度矩阵和聚类中心矩阵。

$$u_{ij} = \frac{1}{1 + (d_{ij}^2 / \eta_j)^{1/(\beta-1)}} \quad (1.11)$$

关于聚类中心的矩阵 $\mathbf{V}$ 仍然和 FCM 算法中式 (1.8) 一样。从式 (1.11) 中可以看出，目标 i 到类别 j 的隶属度仅和目标到该类聚类中心的距离有关。这个特性可以使 PCM 算法检测到噪声点。然而，一些学者指出 PCM 算法这种独立地对待各类的做法有时会产生不理想的聚类结果^[51]。实际上，PCM 算法容易使生成的不