基于规则的建模方法的可解释性及其发展

周志杰[1] 曹 友[1] 胡昌华[1] 唐帅文[1] 张春潮[1] 王 杰[1]

摘 要 建模方法的可解释性指其以可理解的方式表达实际系统行为的能力. 随着实践中对可靠性需求的不断提高, 建立出可靠且可解释的模型以增强人对实际系统的认知成为了建模的重要目标. 基于规则的建模方法可更直观地描述系统机理, 并能有效融合定量信息和定性知识实现不确定信息的灵活处理, 具有较强的建模性能. 本文从基于规则的建模方法出发, 围绕知识库、推理机和模型优化梳理了其在可解释性方面的研究, 最后进行了简要的评述和展望.

关键词 基于规则的建模方法, 可解释性, 系统建模, 不确定性

引用格式 周志杰, 曹友, 胡昌华, 唐帅文, 张春潮, 王杰. 基于规则的建模方法的可解释性及其发展. 自动化学报, 2021, **47**(6): 1201−1216

DOI 10.16383/j.aas.c200402

The Interpretability of Rule-based Modeling Approach and Its Development

ZHOU Zhi-Jie[1] CAO You[1] HU Chang-Hua[1] TANG Shuai-Wen[1] ZHANG Chun-Chao[1] WANG Jie[1]

Abstract The model interpretability refers to the ability to express the real system behavior in an understandable way. With the increasing of reliability requirements in engineering practice, establishing a reliable and interpretable model to enhance human understanding of real systems has become one of the main objectives. Rule-based modeling approach can describe the system mechanism more intuitively. It can not only effectively integrate quantitative information and qualitative knowledge, but can also deal with uncertain information flexibly. This paper combs researches on the interpretability of rule-based modeling approach around the knowledge base, inference engine and model optimization, and finally makes a brief review and prospect.

Key words Rule-based modeling approach, interpretability, system modeling, uncertainty

Citation Zhou Zhi-Jie, Cao You, Hu Chang-Hua, Tang Shuai-Wen, Zhang Chun-Chao, Wang Jie. The interpretability of rule-based modeling approach and its development. *Acta Automatica Sinica*, 2021, **47**(6): 1201−1216

系统建模是一种简化实际工程系统的技术, 通过它可以对实际系统进行分析、预测、仿真和改进, 以满足工程应用需求[1]. 近年来, 以深度神经网络 (Deep neural networks, DNN) 为代表的黑箱模型由于其良好的操作性与建模精度, 在诸如人脸识别[2]、智慧医疗[3] 和声纹识别[4−5] 等领域中得到了广泛的应用. 然而, 黑箱模型的内部参数与结构难以被人理解, 其输出的合理性与可靠性也难以得到验证, 这从一定程度上增加了黑箱模型在实际工程应用中的潜在风险, 尤其是在如医疗诊断决策等对安全性敏感的任务中[1, 6−7]. 在工程实践中, 为了满足实际系统的高可靠性需求, 一方面是在实际系统的可靠性设计中考虑更多因素. 另一方面, 就是通过合理的建模方法建立出可靠且可理解的模型以增强人类对实际工程系统的认识, 获得系统的真实状态, 并据此采取合适的策略对实际系统进行改进和维护, 确保其安全可靠地工作[1]. 因此, 建模方法的可解释性逐渐成为了系统建模必须考虑的重要因素.

建模方法的可解释性指其以可理解的方式表达实际系统行为的能力, 它有利于提高模型的可信度. 在构建模型时, 模型的结构与表达方式清晰易理解, 能融入真实系统的设计原理与经验知识; 在推理计算时, 保持过程的合理与透明性; 在模型优化时, 保证优化后模型的上述性质不被破坏. 这些都利于建立人与模型间的信任关系, 并进一步增强人与模型之间的交互与协作[1, 8−9]. 为了提高系统建模方法的可解释性, 诸多研究人员就模型的可解释性问题开展了广泛和深入的研究, 但由于对可解释性的理解存在主观性, 目前学术界还未给出其统一的定义[1, 8−12]. 在相关研究中, 不同研究者看问题的角度不同, 赋予了 “可解释性” 不同的含义, 因而所提出的可解

收稿日期 2020-06-10 录用日期 2020-09-22
Manuscript received June 10, 2020; accepted September 22, 2020
国家自然科学基金 (61773388, 61751304, 61833016, 61702142), 陕西省杰出青年基金 (2020JC-34) 资助
Supported by National Natural Science Foundation of China (61773388, 61751304, 61833016, 61702142), Shaanxi Outstanding Youth Science Foundation (2020JC-34)
本文责任编委 李鸿一
Recommended by Associate Editor LI Hong-Yi
1. 火箭军工程大学导弹工程学院 西安 710025
1. Missile Engineering College, Rocket Force University of Engineering, Xi'an 710025

释性方法也各有侧重.

模型的可解释性可以分为两类: 1) 事后可解释性 (Post-hoc 可解释性), 旨在通过设计高保真的解释方法或者构建高精度的解释模型对原本可解释性弱的模型进行解释, 以可理解的方式展现其工作机制[13]. 例如, 随机森林 (Random forest, RF)[14], 提升树 (Boosting tree, BT)[15], DNN[2−5, 16] 等模型由于参数量大, 结构复杂, 工作机制透明性低, 属于一类可解释性弱的模型. 针对这些模型, 研究人员提出包括规则提取[17−20]、模型蒸馏[21]、激活最大化[22]、敏感性分析[23−25]、特征反演[26]、类激活映射[27] 等在内的一系列方法帮助人们从整体或者局部理解模型内部的复杂机理[28]. 2) 事前可解释性 (Ante-hoc 可解释性), 指模型本身内置的可解释性, 即无需采取额外的手段便可理解模型的工作机制. 在建模中为了实现 Ante-hoc 可解释性, 通常选择结构透明、易于理解的自解释模型, 例如线性回归[29]、朴素贝叶斯[30]、决策树[31]、基于规则的模型[32] 等. 对于线性模型, 可以通过模型权重体现特征之间的相关关系, 并通过矩阵计算线性组合样本的特征值, 最终复现线性模型的决策过程[33]. 对于朴素贝叶斯模型, 其概率推理过程是透明可理解的[34]; 而决策树模型中, 每一条从父节点到子节点的路径都可转化为一条二分产生式规则, 形成可追溯的推理过程[35−36]. 但是单个决策树模型的建模能力较弱, 在实际工程中往往集成多个决策树构成集成学习相关方法, 其可解释性也随之降低[37]. 由于规则能够更加直观地描述和解释系统机理, 因而基于规则的建模方法 (Rule-based modeling approach, RBM) 在理论及应用中得到了更加广泛的认可与关注.

人工智能领域的符号主义学派强调通过逻辑来描述和模拟人的智能行为, 将知识表示系统分为四类: 1) 基于逻辑的知识表示; 2) 语义网络; 3) 基于框架的知识表示; 4) 产生式知识表示. 产生式知识表示最早来源于美国数学家 Post 提出的波斯特计算模型 (Post machine), 它描述的是人类思维判断中的一种固定逻辑结构关系, 即常见的 “原因−结果”、“条件−结论”以及 “前提−操作”等[38]. 产生式知识表示有多种可以互相转化的形式, 例如逻辑蕴含式和产生式规则. 在现有的研究中, 基于规则的建模方法中的 “规则” 指的就是产生式规则. 产生式规则描述为 “IF-THEN” 形式, 从 Newell 和 Simon 等开发的基于规则的产生式系统开始, 已逐渐成为了专家系统、人工智能等领域中应用最广泛的知识表示方法[1, 32]. 最初的产生式规则对确定性信息具有较为直观的描述能力. 1965 年, Zadeh 在 *Information and Control* 杂志发表 “Fuzzy sets”, 提出了一种利用模糊集合和模糊逻辑分析复杂系统的新方法[38]. 由该方法衍生的模糊规则, 如 Mamdani-Larsen (ML) 型规则[1, 39] 及 Takagi-Sugeno (TS) 型规则[1, 40−41], 通过语义变量和模糊命题为系统行为的描述提供了更加易于理解的表达形式, 且更加符合人类的推理模式[32]. 在诸多领域中, 基于模糊规则的建模方法能够有效利用定量信息和定性知识处理模糊不确定性. 但是在实际工程系统中, 模糊不确定性 (Fuzzy uncertainty)、不完备性 (Incompleteness)、概率不确定性 (Probabilistic uncertainty) 等各类不确定性广泛共存, 这严重制约了模糊规则的建模性能[42−45]. 为此, Yang 等于 2006 年在 Dempster-Shafer 证据理论[46−47]、决策理论[48]、模糊理论[38] 和传统产生式规则[1, 32, 38] 的基础上, 将置信框架引入传统产生式规则, 提出了基于证据推理算法的置信规则库推理方法 (Belief rule-base inference methodology using the evidential reasoning approach, RIMER)[42, 45]. RIMER 方法的核心之一是置信规则库 (Belief rule base, BRB), 它是传统模糊规则库 (Fuzzy rule base, FRB) 的一般化, 能为工程系统中知识提供更加可靠的描述[45].

RBM 的优势在于其内置的可解释性, 本文的 “可解释性” 侧重于说明如何建立和保持 RBM 的内置可解释性, 即如何在构建知识库时获得可解释性, 如何保证推理过程的透明可解释, 如何在优化过程中保持可解释性. RBM 的可解释性主要取决于几个因素, 包括模型结构、规则数目、模糊集的形状等[8, 10, 45]. Zhou 等结合前人的研究从模糊集和模糊规则层面将 FRB 的可解释性分为低级可解释性 (Low-level interpretability) 和高级可解释性 (High-level interpretability)[10]. 然而, 随着模糊规则被拓展为置信规则, 高低级可解释性在描述 BRB 系统的可解释性方面存在一定的局限性. 本文主要对 RBM 的发展进行概述, 从知识库、推理机、模型优化等角度分析总结典型文献中关于其可解释性的研究, 并在此基础上对其未来发展进行了阐述, 旨在为 RBM 的发展和改进提供一定的参考和借鉴. 本文的整体框架如图 1 所示.

1 RBM 基础理论简介

构建 RBM 时, 一般先从专家或领域知识获取规则以构成一个知识库, 随后设计有效的推理引擎实现对观测信息的推理从而得到结论[42, 45]. RBM 中产生式规则的一般形式为:

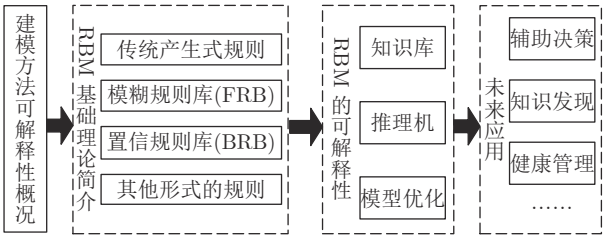

图 1　论文整体框架

Fig. 1　The overall framework of the paper

$$R = \langle \boldsymbol{X}, \boldsymbol{A}, \boldsymbol{D}, F \rangle \tag{1}$$

其中, $\boldsymbol{X} = \{X_i; i = 1, \cdots, T\}$ 是属性的集合, 且每个属性都可从有限集合 $\boldsymbol{A} = \{A_1, A_2, \cdots, A_T\}$ 中获取值 (或命题). $A_i = \{A_{ij}; j = 1, \cdots, J_i = |A_i|\}$ 表示由 X_i 的参考值 (或假设) 组成的集合, 这些参考值可以是定性的也可以是定量的; $\{X_1 \to A_1, X_2 \to A_2, \cdots, X_M \to A_M\}$ 定义了一系列的前提条件, 它们代表了所研究问题的基本状态, 可由逻辑 "与" 或者逻辑 "或" 连接; $\boldsymbol{D} = \{D_j; j = 1, \cdots, N\}$ 表示由结果组成的集合, 且 D_j 可以表示一个结论或一个动作; F 表示一个逻辑函数, 反映了前提条件与结果 $\boldsymbol{D}$ 之间的关系.

产生式规则自 1934 年提出以来, 其发展主要围绕不确定性知识的描述展开[49]. 传统的产生式规则一般用于对确定性信息的描述. 但是随着人们对工程系统认识的不断加深, 越来越多的研究认为处理不确定性信息的能力对提高模型可解释性与建模精度有着重要意义. 基于此, 传统的产生式规则也逐渐扩展成为模糊规则和置信规则[50−51]. 本节将从不确定性信息处理能力角度对 RBM 的发展进行介绍. 同时, 本节也将对决策树和贝叶斯网络等一类可视为特殊规则的模型进行介绍, 以帮助读者更加全面理解 RBM 的发展.

1.1　基于传统产生式规则的建模方法

传统产生式规则通常可以写成如下形式:

$$\text{IF } X \text{ THEN } Y \tag{2a}$$

其中, X 为产生式规则的前提 (前件), 它一般由事实的逻辑组合构成, 表示该规则被使用的条件; Y 表示一组结论 (后件). 产生式规则的含义是如果前提 X 满足, 则可得到结论 Y.

下面以一条医生判断病人是否感冒的规则为例对传统产生式规则进行解释:

$$\text{IF 发烧} \wedge \text{咳嗽 THEN 病人感冒} \tag{2b}$$

其中事实的逻辑组合 "发烧 $\wedge$ 咳嗽"是该条规则是否被使用的先决条件, $\wedge$ 为逻辑连接符, 表示逻辑 "与". 当上述事实的逻辑组合满足时, 该条规则被激活产生结论 "病人感冒". 由这个例子可以看出, 产生式规则能简单明确地描述专家知识, 符合人类推理的思维模式. 但是上述规则仅能够确定性的描述 "发烧"、"咳嗽"和 "感冒", 并不能进一步的解释病人发烧/咳嗽/感冒的程度.

1.2　基于 FRB 的建模方法

基于 FRB 的建模方法可分为两类[1], 即精确模糊建模方法 (Precise fuzzy modeling, PFM) 和语义模糊建模方法 (Linguistic fuzzy modeling, LFM). TS 型模糊规则与 ML 型模糊规则是构成 FRB 的两种常用组件. 其中, 基于 TS 型规则库的建模方法是 PFM 的代表方法之一, 其主要目的是获得具有良好精度的模糊模型, 这种方法在模糊变量的使用上一般不会赋予模糊集相关意义. 基于 ML 型规则库的建模方法是 LFM 的代表方法之一, 主要目的是获得具有良好可解释性的模糊模型, 在其规则中, 前件和后件使用由语义术语组成的语言变量和模糊集来定义相关意义. 同理, 与 PFM 相比, 在推理机的选择与构造上, LFM 也更倾向于保证推理过程的透明和推理结果的可解释性.

TS 型规则库中第 k 条规则可描述如下[1, 40−41]:

$$\begin{aligned} R_k : \text{IF } & X_1 \text{ is } A_1^k \wedge X_2 \text{ is } A_2^k \wedge \cdots \wedge X_{T_k} \text{ is } A_{T_k}^k \\ & \text{THEN } Y^k(\boldsymbol{X}) = c_0^k + c_1^k X_1 + \cdots + c_{T_k}^k X_{T_k}, \\ & k = 1, \cdots, L \end{aligned} \tag{3}$$

其中, L 为规则条数, $\wedge$ 为连接符, 表示逻辑 "与", 结合实际需求, 逻辑连接符也可以是表示逻辑 "或" 的 $\vee$; A_i^k 表示第 k 条规则中第 i 个输入的参考值, 一般为一组带有物理意义且相互排斥的语义值; Y^k 为输出结果, c^k 为输出结果的参数. TS 型模糊规则将仿射函数作为模糊规则的输出结果, 通过将复杂非线性系统局部分解为一系列线性系统, 使其具备较好的逼近性能. TS 型模糊规则在控制领域得到了广泛的应用[52−53].

与 TS 型规则相比, ML 型规则的后件具有更加简洁直观的表达方式, ML 型规则库第 k 条规则描述为[1, 39]:

$$\begin{aligned} R_k : \text{IF } & X_1 \text{ is } A_1^k \wedge X_2 \text{ is } A_2^k \wedge \cdots \wedge X_{T_k} \text{ is } A_{T_k}^k, \\ & \text{THEN } D_n^k,\ n = 1, \cdots, N, k = 1, \cdots, L \end{aligned} \tag{4}$$

其中, D_n^k 为表述语义的输出结果, 语义值也可对应设置数值参考值, 转化为数值输出.

由式 (3) 与 (4) 可以看出, ML 型规则和 TS 型规则具有相同的前提结构和不同的后件部分. ML

型规则的后件部分是一个模糊集, 用以更加直观地描述实际系统. TS 型规则的是一个线性仿射函数, 用以精确地拟合实际系统. 在建立 ML 型规则时, 其关键在于将定性与定量知识转化为由语义术语组成的语言变量和模糊集. 在建立 TS 型规则时, 其关键在于选择合适的函数对实际系统进行拟合.

现实系统中, 存在着多种不确定性. 举例来说, "约翰很年轻的可能性是 0.8". "年轻"一词是年龄的模糊表达, 带有模糊不确定性. "0.8"是我们给出该信息的确定度, 带有概率不确定性. ML 型模糊规则能较好地描述和处理模糊不确定性, 但是难以较好处理广泛存在的概率不确定性[42, 45]. 置信框架是一种描述信息不确定性的理想方法[42−45], 因此有必要结合置信框架对模糊规则进行一般化的扩展.

1.3 基于 BRB 的建模方法

BRB 通常由一系列置信规则构成, 其中第 k 条规则表述为[42, 45]:

$$\begin{aligned} R_k: &\text{ IF } X_1 \text{ is } A_1^k \wedge X_2 \text{ is } A_2^k \wedge \cdots \wedge X_{T_k} \text{ is } A_{T_k}^k, \\ &\text{THEN } \{(D_1, \beta_{1k}), \cdots, (D_N, \beta_{Nk})\}, (\sum_{n=1}^{N} \beta_{nk} \le 1), \\ &\text{with rule weight } \theta_k \ \ (k = 1, \cdots, L) \text{ and} \\ &\text{attribute weights } \delta_i \ \ (i = 1, \cdots, T_k) \end{aligned} \tag{5}$$

其中 A_i^k $(i = 1, \cdots, T_k)$ 表示第 k 条规则中第 i 个前提属性 X_i $(i = 1, \cdots, T_k)$ 的参考值; T_k 表示第 k 条规则中前提属性的数量; $\wedge$ 表示逻辑关系 "与"; θ_k 表示第 k 条规则相对于其他规则的重要度, 称为规则权重; δ_i $(i = 1, 2, \cdots, T_k)$ 表示第 i 个前提属性 X_i $(i = 1, \cdots, T_k)$ 相对于其他属性的重要度, 称为属性权重; $0 \le \beta_{nk} \le 1$, $(n = 1, 2, \cdots, N)$ 表示第 k 条规则中对于第 n 个结果 D_n 的支持度, 也称对 D_n 的置信度; 上述置信规则的完整性取决于其规则后件所描述的置信分布是否存在无知, 即是否存在未分配给辨识框架中任意结果 D_n 的置信度. 当 $\sum_{i=1}^{N} \beta_{ik} = 1$ 时, 第 k 条规则后件描述的置信分布不存在未分配的置信度, 可认为第 k 条规则是完整的, 否则, 第 k 条规则是不完整的.

在后续研究中, Chang 等对上述置信规则的假设进行了扩展, 提出了基于 "或" 假设的置信规则及相应的激活权重计算方法, 并进一步探讨了两种假设下置信规则的相互关系[54−57]; Liu 等认为上述置信规则的前件部分不能较好地描述不确定性, 于是对置信规则的前件进行了扩展, 提出了扩展置信规则库模型 (Extended BRB model)[58−59].

1.4 其他特殊形式的规则

产生式规则在描述人类知识方面具有较好的通用性, 例如, 图搜索中的状态转换规则、程序设计的词法规则、逻辑中的逻辑蕴含式、数学中的微积分公式、化学中分子式的分解变换规则等都可用产生式规则来描述. 现有研究中, 许多建模方法中的知识表示方法均可以轻松地转化为产生式规则, 例如树模型 (决策树、故障树) 和贝叶斯网络等. 因此, 这些知识表示方法也可以视作特殊形式的产生式规则. 下面将以故障树以及贝叶斯网络为例, 对其与产生式规则的转化关系进行介绍.

1) 故障树与产生式规则的转化

故障树是一种因果演绎分析方法, 它把系统的故障 (结果) 及其发生的直接原因作为顶事件和底事件, 采用逻辑门来描述故障原因间的逻辑关系, 用树形结构来描述原因与结果之间的因果关系. 可以看出, 故障树的知识表示方式与产生式规则具有极强的关联性, 两者本质上的逻辑关系是等价[60].

图 2 所示的是故障树中常见的两个基本单元, 其与产生式规则的对应转化关系为: 故障树中的 "与门" 表明多个基本事件同时发生时顶事件才会发生, 这与产生式规则中 "交" 逻辑一致. 因此, "与门" 故障树转化的产生式规则为:

$$\begin{aligned} &\text{IF } (X_1 \text{ is } Fault) \wedge \cdots \wedge (X_T \text{ is } Fault), \\ &\text{THEN } \{(D_{Normal}, 0), (D_{Fault}, 1)\} \end{aligned} \tag{6a}$$

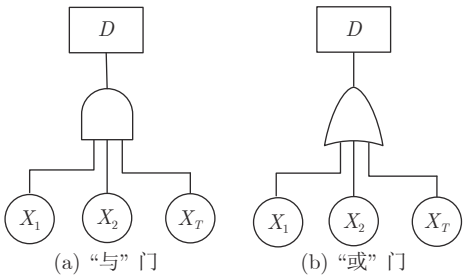

图 2 故障树

Fig. 2 Fault tree

同理, 故障树中的 "或门" 表明多个基本事件中, 只要有一个事件发生, 顶事件就会发生, 这与产生式规则中 "并" 逻辑一致. 因此, "或门" 故障树转化的产生式规则为:

$$\begin{aligned} &\text{IF } (X_1 \text{ is } Fault) \vee \cdots \vee (X_T \text{ is } Fault), \\ &\text{THEN } \{(D_{Normal}, 0), (D_{Fault}, 1)\} \end{aligned} \tag{6b}$$

2) 贝叶斯网络与产生式规则的转化

贝叶斯网络是一种基于有向无环图模型来描述因果关系的概率模型. 通常, 一个复杂的贝叶斯网

络可以分解为若干个如图 3 所示的贝叶斯网络片段. 每个贝叶斯网络片段包含一个子节点和若干个父节点, 节点由有向弧连接, 从父节点指向子节点. 贝叶斯网络将父节点与子节点分别视作前提条件和推理结论, 并通过概率表来描述父节点与子节点间的关系.

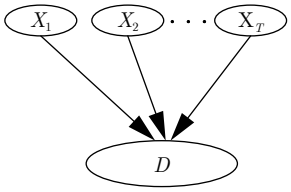

图 3　贝叶斯网络片段

Fig. 3　Bayesian network fragment

在如图 3 所示的贝叶斯网络片段中, 每个父节点 X_i 可能含有多个状态 $A_i=\{A_{i,1},\cdots,A_{i,J}\}$ $(i=1,\cdots,T, j=1,\cdots,J)$, 子节点 D 也可能有多个状态 $D_1,\cdots,D_N$. 因此, 该贝叶斯网络片段可转换为如下产生式规则 (置信规则):

$$\begin{aligned}&\text{IF } X_1 \text{ is } A_1 \wedge \cdots \wedge X_T \text{ is } A_T,\\&\text{THEN } \{(D_1,\beta_1),\cdots,(D_N,\beta_N)\}\\&\text{with rule weight } \theta \text{ and}\\&\text{attribute weights } \delta_i \ \ (i=1,\cdots,T)\end{aligned} \tag{7}$$

其中 $\beta_1,\cdots,\beta_N$ 为相对于子节点各状态的置信度, θ 和 δ 分别表示规则权重和属性权重. 这些参数均可以通过分析概率表和专家经验来确定[61].

2　RBM 的可解释性

可解释性是 RBM 的固有特征[10, 62], 现有研究主要从三个方面对其展开探讨: 知识库 (Knowledge base)、推理机 (Reasoning engine) 和模型优化 (Model optimization). 此外, 一些文献也对该方法的 Post-hoc 可解释性进行了研究. 本节主要内容如图 4 所示.

2.1　知识库

在 RBM 中, 知识库由一系列规则组成. 可解释的知识库需要具备清晰明确的语义, 完备、简洁、一致的规则以及具有物理意义的结构和参数[1].

2.1.1　清晰明确的语义

1) 语义的可区分性

语义可区分性用来描述规则输入划分空间的合理性. IF-THEN 规则输入的参考值及其匹配区间应该具备可区分性, 以表示一个明确的语义[10, 63–64].

2) 匹配度的标准化与互补性

匹配度的标准化指论域 U 内至少存在一个数据点相对于某个参考值的匹配度等于 1, 且所有参考值的匹配度应该处于 0 到 1 之间, 即[10, 64]:

$$\begin{aligned}&\forall\, 1\le\xi\le T,\ \ \exists\ x_0\in U, a_\xi(x_0)=1\\&\forall\, 1\le\xi\le T,\ \ x\in U, 0\le a_\xi(x)\le 1\end{aligned} \tag{8}$$

其中, T 表示前提属性参考值的数量, x_0 表示论域 U 内的某个定值. $a_\xi(\bullet)$ 表示相对于第 ξ 个参考值的匹配度.

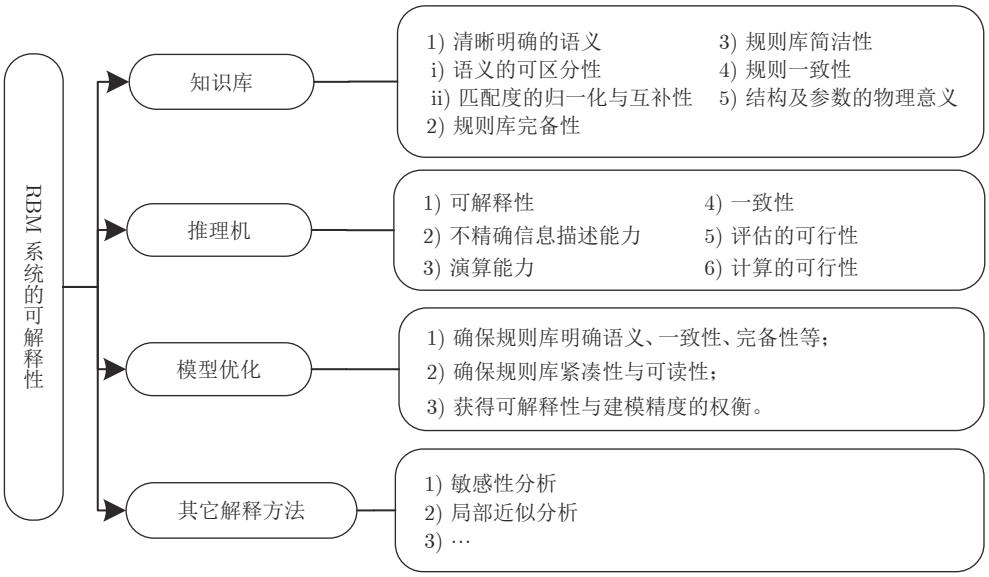

图 4　RBM 可解释性的主要内容

Fig. 4　The main contents of RBM interpretability

匹配度的互补性指在论域 U 内前提属性空间参考值的所有匹配度之和应该小于或等于 1, 即[10]:

$$\forall\, 1\leq\xi\leq T \text{ and } x\in U,\ \ 0\leq a_{\xi}(x)\leq 1, \sum_{\xi=1}^{T} a_{\xi}(x_0)\leq 1 \tag{9}$$

其中, 当 $\sum_{\xi=1}^{T} a_{\xi}(x_0)=1$ 时认为输入信息是完整的, 否则认为输入信息有缺失. 匹配度的标准化与互补性在实际工程中有助于形成易于理解的语义描述.

要保证明确的语义, 一种方法是通过专家经验确定相应的前提属性的参考值与匹配区间. 另一种常用的方法是通过使用模糊集之间的相似性测度来合并相似的模糊集以保证其可区分性. 常用的相似性测度为 Jaccard 相似系数 (Jaccard similarity coefficient), 其计算方法如下所示[65]:

$$\text{sim}(A,B)=\frac{|A\cap B|}{|A|+|B|-|A\cap B|} \tag{10}$$

其中 A 和 B 表示两个模糊集, $\cap$ 和 $|\bullet|$ 分别表示集合的交集和基数. Jaccard 相似系数的范围为 0 到 1. 当两个模糊集完全相同时, Jaccard 相似系数的值为 1; 当二者完全不相似时, Jaccard 相似系数的值为 0. Zwick 等对模糊集之间的许多相似性测度进行了比较分析[66].

此外, Mencar 等还提出了一种基于可能性测度的量化方法[67−68], 该方法通过量化模糊集重叠的程度来评估可区分性, 并且可以有效减少相似模糊集合合并过程中的计算量, 其定义如下:

$$\Pi(A,B)=\sup_{x}\left(\min\left(\mu_A(x),\mu_B(x)\right)\right) \tag{11}$$

其中 A 和 B 表示两个模糊集, $\mu_A(x)$ 表示输入 x 对于模糊集 A 的隶属度. 现有用于量化模糊集可分辨性的方法大多只基于输入数据, 而没有使用输出数据中包含的信息. Zhou 和 Gan 提出了一种局部熵, 通过考虑输入输出样本提供的信息来实现模糊集可区分性的测量[69].

2.1.2 规则库的完备性

规则库的完备性指对于任何一种可能的输入 (定量或定性) 都应该至少匹配一个参考值, 并且至少应激活一个规则, 即一个可解释的规则库应该包含系统的所有工作模态, 其描述如下[70]:

$$\forall x\in U,\ \exists\, 1\leq\xi\leq T,\ a_{\xi}(x)>0 \tag{12}$$

$$\forall\, x\in U,\ \ \exists 1\leq l\leq L, 0<w_l\leq 1 \tag{13}$$

其中 T 表示前提属性参考值的数量. U 是 x 的整个可行域. $a_{\xi}(x)$ 表示与第 ξ 个参考值的匹配程度. L 表示规则数. w_l 表示第 l 条规则的激活权重.

Meesad 等认为完备性主要包括两个要素: 模糊划分的完备性和模糊规则结构的完备性, 并分别提出了相应的完备性测度[71]. Espinosa 等提出了具有语言完整性的自治模糊规则提取算法, 在保持规则库语义完整性及覆盖性的同时, 能够很好地拟合输入输出数据[72]. Li 等通过模糊减法聚类 (Fuzzy subtraction clustering) 计算出输入数据的最佳聚类数和聚类中心, 并将其作为各属性参考值, 该方法在输入信息完备的前提下能有效保证规则库的完备性[73]. 为了保证参考值在优化过程中对空间分区覆盖的可解释性, De Oliveira 等在优化过程中引入了覆盖约束[64]. 不同于上述研究, Zhou 等提出了基于 “效用统计” 的规则库在线构造方法, 该方法能依据输入模态的变化实时地增加和删除规则, 保证规则库的紧凑和完备[74].

2.1.3 规则库的简洁性

规则库的简洁性是保证其可解释性的关键要求之一[75]. 根据 Occam 剃须刀原理 “The best model is the simplest one fitting the system behaviors well”, 在保证模型建模性能处于满意水平的前提下, 规则库的规模必须尽可能小以易于对工程系统的全局性理解. 规则库中规则的数量会随着前提属性和参考值的数量增长呈指数增长, 即规则的组合爆炸[42, 45]. 这严重限制了规则库模型的工程应用. 因此, 必须采取一定的手段保证适当数量的前提属性和参考值.

特征选择的目标是在建模精度和模型复杂度之间找到最优的平衡点, 产生一个紧凑、简洁、具有良好泛化和解释能力的规则模型[1]. 常用的特征选择算法有过滤式 (Filter) 特征选择算法与包装器 (Wrapper) 特征选择算法[1, 76]. Filter 特征选择算法按照特征的某种特性 (例如相关性) 对各个特征进行评分, 设定阈值进行特征筛选. 该方法简单有效, 但是由于没有启发式学习过程的参与, 往往不能得到最佳特征子集. Wrapper 特征选择算法需要建立相应的目标函数, 并利用学习算法获得模型的精度估计, 以选择最优特征子集. 但是该方法的主要问题是当特征数量比较大的时候, 其效率低下.

规则库模型结构学习的主要目的是通过启发式学习方法确定合适数量的属性参考值, 从而降低模型复杂度. 例如, Chang 等首先利用多维约简技术 (Multiple dimensionality reduction techniques) 提出了 BRB 结构学习方法[77−78]. 王应明等通过对粗糙集理论的推导进一步探索了 BRB 结构学习[79]. Chang 等通过将结构和参数学习与 Akaike 信息准

则 (AIC) 结合构建优化目标函数, 有效降低了模型复杂度[80]. 当工程系统较为复杂时, 构建合理结构的层次模型, 使其规则数量随前提属性的数量线性增加, 有利于降低模型复杂度[45].

此外, Chang 等还提出了一般化的析取假设下的 BRB 模型 (Disjunctive BRB), 并在有限范围内讨论了其与传统 BRB 模型间的转化关系与一致性[55]. 同时, 也提出了一种在传统假设下的 BRB 结构和参数优化方法, 并将其扩展到析取假设下, 为降低规则库规模提供了新的思路[57].

2.1.4　规则的一致性

规则的一致性能有效防止在推理过程及输出结果中出现歧义. 对于可解释的规则库, 相互冲突的规则是不能共存的. 如果规则库是由专家知识提取得到, 规则的一致性容易得到保证. 但数据驱动的规则提取过程往往会因为带噪声的数据而产生冲突规则. 严重不一致的规则无疑会导致规则库性能退化, 并使之无法解释. Jin 等认为规则的一致性不应局限于规则之间, 还应该考虑规则与人类常识之间的一致性, 并给出了规则的一致性定义如下[75]:

$$\mathrm{Cons}\left(R_i,R_k\right)=\exp\left\{-\frac{\left(\dfrac{SRP\left(R_i,R_k\right)}{SRC\left(R_i,R_k\right)}-1\right)^2}{\left(\dfrac{1}{SRP\left(R_i,R_k\right)}\right)^2}\right\} \tag{14a}$$

其中 R_i,R_k 代表第 i 条规则与第 k 条规则, $SRC(\bullet)$ 表示规则后件的相似性, 定义为

$$SRC\left(R_i,R_k\right)=S\left(B^i,B^k\right) \tag{14b}$$

$S(\bullet,\bullet)$ 是规则后件模糊集 B^i 和 B^k 之间的相似性度量. 在置信规则中, 其表示为后件置信分布之间的相似性度量. $SRP(\bullet)$ 表示规则前提的相似性, 定义为

$$SRP\left(R_i,R_k\right)=\min_{j=1}^{n}S\left(A_j^i,A_j^k\right) \tag{14c}$$

$SRP(\bullet)$ 表示规则前提的相似性, A_j^i,A_j^k 表示规则前件模糊集.

上述定义中, 在两条规则的 SRP 和 SRC 成正比例的情况下, 如果两个规则的 SRP 较高, 则其一致性程度会趋向于低. 在特殊情况下, 当两条规则具有相同的前提和结果, 或两条规则的 SRP 和 SRC 相等, 或两条规则的 SRP 等于 0 时, 一致性程度达到最大值 1. 但是当两条规则具有相同的前提和结果时, 往往会出现冗余, 影响模型的紧凑性. 此外, 如果规则前提相同但规则后件不同, 两条规则的一致性范围为 0 到 1. 如果两条规则的前提非常不同 (SRP 非常低), 其一致性程度则处于较高水平. 这符合规则的一致性假设, 即两条带有较大差异前提的规则常被认为是一致的[10, 75, 81].

2.1.5　结构及参数的物理意义

带有物理意义的结构主要体现在两方面: 规则的结构和规则库的结构. 对于规则的结构, ML 型规则和 TS 型规则是两种应用最为广泛的模糊规则, 二者具有相同的前提结构和不同的后件部分. ML 型规则的后件部分是一个模糊集, TS 型规则的是一个线性仿射函数. 相比较于 TS 型规则, ML 型规则的模糊集在描述知识方面更加直观. 置信规则就是在 ML 型规则的基础上扩展而来. 对于规则库的结构, 为了清晰描述复杂工程系统内部各组件间的逻辑关系, 规则库一般会基于层次结构来构建, 以明确输入间的传递关系. 在推理时, 采用自底向上的推理方法, 逐层对证据进行聚合产生更高级别的证据[42, 45]. 例如, Yang 等和 Hodges 等构建的用来确定容器包含石墨的置信度的层次模糊规则库, 其中每一条规则及组件间均有较强的逻辑关系[45, 82].

在规则的参数方面, 以置信规则为例, 其参数主要包括属性权重、规则权重、激活权重以及置信度等. 其中属性权重 $\delta\in[0,1]$ 表示该前提属性相对于其他属性的重要度; 规则权重 $\theta\in[0,1]$ 表示该条规则相对于其他规则的重要度; 激活权重 $w\in[0,1]$ 表示对应规则被输入激活的程度, 它由输入信息、属性权重与规则权重共同决定; 置信度 $\beta_i\in[0,1]$ 表示对结果的支持程度. 上述参数具有一定的主观性, 其值的大小一般依赖于专家经验. 此外, Feng 等[24] 还在 BRB 中引入了属性可靠度 $r\in[0,1]$ 用以表示输入信息的可靠程度. 属性可靠度体现了属性的客观性, 它与系统的内部机理、工作环境密切相关.

2.2　推理机

知识库的可解释性保证了其能以可理解的方式准确描述实际系统中的不确定信息. 为了产生可靠的结果, 推理机对不确定信息的处理能力至关重要. 澳大利亚西澳大学 Walley 教授认为可从六个标准来对现有不确定性推理方法进行评价[83−84]:

标准 1. 可解释性 (Interpretation). 指推理方法对不确定信息的处理应有足够明确的解释, 可以用于指导评估, 理解结论, 并将其作为行动的基础.

标准 2. 不精确信息描述能力 (Imprecision). 指推理方法对不精确信息的描述能力, 即能够对部分或全部的无知、有限的或相互矛盾的信息以及不确定性的不精确评估.

标准 3. 演算能力 (Calculus). 指推理方法应该

具有整合不确定性信息的能力, 即能够结合新的信息更新不确定度, 并使用它们来计算其他不确定度, 最终得出结论并做出决策.

标准 4. 一致性 (Consistency). 指推理方法的不确定性推理过程应该和假设保持一致.

标准 5. 评估的可行性 (Assessment). 指推理方法需要具备处理各种类型信息的能力, 包括用自然语言表示的不确定性, 例如 "if A then probably B". 同时也能将定性判断与定量的不确定性信息结合起来.

标准 6. 计算的可行性 (Computation). 指推理方法能够通过合理高效的计算过程得到推理结论.

上述六个标准从理论和实践对不确定性推理方法进行了衡量, 其中标准 1 (Interpretation) 是对不确定性推理方法的基本要求, 它可以支撑演算过程, 保证模型的一致性, 同时还能够指导评估过程[83−85].

常用的不确定性推理方法主要包括贝叶斯概率推理方法 (Bayesian probability reasoning)、信任函数 (Belief function)、可能性推理方法 (Possibility reasoning) 以及似然推理方法 (Likelihood reasoning). 本文将结合上述六种标准对上述不确定性方法进行简要分析.

1) 贝叶斯概率推理方法

贝叶斯概率方法通过无条件概率 $P(A)$ 或者条件概率 $P(A|B)$ 对不确定性进行描述, 具有简单的行为解释能力, 这使得其不确定性整合过程变得合理, 也保证了推理结果的一致性. 贝叶斯概率推理依赖于精确的概率描述, 难以有效模拟实际工程系统中广泛存在的无知性、信息不完备性, 也无法处理自然语言中的不确定性与专家意见之间的冲突[83, 85].

另一方面, 精确概率的获取本身是一件较为困难的工作, 因此贝叶斯概率推理方法在评估和计算 (标准 5, 6) 方面也存在挑战, 导致其在实际应用中存在诸多限制[86−88]. 早期, 有研究通过简化贝叶斯概率假设, 使用类似于模糊逻辑中 max/min 规则的简单规则来整合证据[89]. 然而, 这会产生不一致的概率. 最近, 有些研究基于专家对变量之间因果关系的理解来判断条件的独立性, 将变量之间的关系以图形化表示形成一些新模型, 例如信念网络、因果网络或有向无环图[87, 90−93]. 这些模型的优势在于有效减少了评估和计算的工作量, 在许多实际问题中这是非常重要的.

总的来说, 贝叶斯概率推理方法在标准 1, 3, 4 上具有较好表现, 但是在标准 2, 5 上不具有优势. 对于标准 6, 贝叶斯概率推理的扩展方法在一定程度上可以满足工程需求.

2) 信任函数理论

信任函数理论是对经典概率论的扩展, 是 Dempster-Shafer (D-S) 证据推理方法的关键理论, 由 Dempster 在 20 世纪 60 年代提出, 后又由 Shafer 不断发展[46−47, 92]. 信任函数是一个定义在辨识框架 Ω 所有子集上的实值函数, 可以描述为如下形式:

$$\underline{P}(A) = \sum_{B \subseteq A} m(B) \tag{15}$$

其中 $m(\bullet)$ 是 Ω 的子集上的概率质量函数. A 和 B 是定义在辨识框架 Ω 上的子集.

依据 Shafer 的阐述, 信任函数可以通过多值映射的精确概率测度生成, 这种多值映射具有一定的行为解释[47]. 但信任函数的演算在很大程度上依赖于 Dempster 的组合规则, 需要有条件独立性的明确判断支持, 因而在实际应用时应更多地注意 Dempster 组合规则的确切使用条件以避免产生直觉上的不一致, 从而影响其解释性和评估过程[93−95]. 信任函数理论在充分考虑 Dempster 组合规则应用条件的前提下, 可在一定程度上较好满足标准 1~4. 另外, 信任函数可以对各类型的局部无知、有限信息或矛盾证据进行建模, 但是难以对语义中的概率判断建模, 难以满足标准 5[96]. 在计算方面, 信任函数理论中的概率质量函数 $m(\bullet)$ 用来描述不确定性信息, 在采用 Dempster 组合规则对概率质量函数进行融合的过程中, 计算的复杂度会随着辨识框架中元素个数增加而呈指数增加, 这就是信任函数面临的 "组合爆炸" 问题. "组合爆炸" 问题较大程度限制了信任函数在实际工程中的应用. 为使信任函数在实践中获得较好的计算效率, 许多研究尝试通过近似的方法来降低计算复杂性, 例如 Voorbraak 定义了信任函数的贝叶斯近似, 该方法能够在不对融合结果产生实质影响的前提下有效减少计算量[97]. 可见, 对于标准 6 而言, 信任函数理论需要进一步结合其他方法来弥补不足.

3) 可能性推理方法

1965 年, Zadeh 教授首次提出了模糊集理论, 通过可能性度量来描述不确定性, 并为自然语言中的模糊表达式提供数学模型以模拟语义的不精确性和模糊性, 有效弥补了人工智能领域中经典概率在处理多种不确定性方面的弱点[38, 98].

一阶可能性测度与二阶可能性测度在可能性推理方法中十分重要[98−99]. 一阶可能性测度用可能性分布来定义, 一般解释为一致的上概率, 具有较好的计算性能, 可以较好满足标准 1, 4, 6. 它可以有效处理语义中的模糊带来的不确定性, 但主要缺陷

是它们不能对许多常见的不确定性类型进行建模, 例如无法模拟自然语言中的精确概率判断[100]. 因此一阶可能性测度难以满足标准 2, 3, 5. 可能性测度是一种非常特殊的上概率, 但上概率本身在许多问题上是不充分的, 一般需要进行上下概率预测[98]. 为此, 需要考虑可能性测度的对偶测度, 即必要性测度. 必要性测度一般解释为一致的下概率. 假设可能性测度表示为 π , 其所描述的上概率可以计算为:

$$\bar{P}(A)=\pi(A)=\sup\{\pi_X(\omega):\omega\in A\} \tag{16}$$

其中 A 为集合 U 的子集. 必要性测度所描述的下概率可以计算为:

$$\underline{P}(A)=1-\sup\{\pi_X(\omega):\omega\in A^c\}=1-\pi(A^c) \tag{17}$$

其中 A^c 表示 A 的补集. 可能性测度与必要性测度满足 $\underline{P}(A)\leq\bar{P}(A)$. 在所有概率分布集合上定义的二阶可能性测度比一阶测度更具表现力, 它能够同时对精确或不精确的不确定性判断进行建模. 但二阶测度比一阶测度复杂得多, 因此很难解释和评估, 例如对定性判断建模来说, 二阶测度就过于复杂[99−100]. 因此, 二阶测度虽然能够弥补一阶测度在标准 2~5 上的不足, 但这是以降低其在标准 1, 5 上的表现为代价的.

4) 似然推理方法

与上述三种方法相比, 似然推理方法在推理不确定性、部分信息无知、自然语言中概率判断的模糊性、专家意见之间的冲突等方面更加具有一般性, 满足标准 1~4[85, 101]. 似然推理方法中上、下概率、期望和条件概率是通过一种自然扩展 (Natural extension) 的技术从初始估计中构造出来的, 具有较好的行为解释, 可以用来验证初始评估的一致性, 并确保评估与结论的一致性[84, 102]. 在复杂问题中应用似然推理进行推论和决策的一般方法是自然扩展, 该方法可以简化为一个线性规划问题. 但还需要进一步的研究来寻找涉及独立性判断的自然扩展有效方法[83−84]. Yang 等提出的 ER 规则可以看做一种特殊的似然概率推理方法[44−45]. Wang 等在此基础上推导出来的 ER 解析算法有效降低了计算量, 使其更具有工程适用性, 但证据不独立的问题仍然困扰着 ER 规则[103].

基于上述四种方法的分析, 这些方法中的每一种都存在一些特殊问题, 似然推理方法更具有普遍意义. 但需要强调的是, 上述分析的目的不是为了判断推理方法的优劣. 事实上, 随着拓展研究的不断推进, 推理方法针对特定问题的改进依然可以较好地满足工程需求.

2.3　模型优化

在工程实践中, 一般通过两种手段实现模型的可解释性[10, 42, 45]. 其一是充分结合实际系统机理信息与专家知识构建满足可解释性要求的模型, 但是这种方法依赖于对实际系统内部机理的完全剖析, 显然这是一项极具挑战的工作. 其二是先利用有限的专家知识构建初始模型, 再利用优化学习方法对初始模型的结构和参数进行调整, 以保证模型可解释性的同时提高其建模性能. 从工程角度来看, 当专家能够完全掌握实际系统机理信息时, 可采用第一种手段构建可解释的模型. 但由于实际工程系统结构复杂, 专家很难完全获得和掌握其机理信息, 相对而言, 有限的专家知识和样本数据是更容易获得的. 因此, 第二种手段在工程中更为普遍. 通过第二种手段构建的模型, 结合了专家从全局角度对真实系统的认识和数据样本中与真实系统紧密相关的实测信息, 这使得模型对真实系统行为的解释能力和工程适用性得到了增强.

在 RBM 中, 面向精度的自适应优化学习算法常常会产生矛盾或者不可辨识的模型参数及结构[45]. 虽然这可能有利于提高模型优化的精度, 但其会严重影响建模方法的泛化性能和可解释性. 建模精度和可解释性是相互制约的目标, 理想的情况是在很大程度上满足这两个标准. 因此, 研究人员通常更期望采用优化算法并根据用户的需求在可解释性和建模精度之间取得最佳的平衡. 为实现这个目标, 需要构建合理的可解释性约束与优化目标函数, 并选择高效的优化学习算法对模型的参数及结构进行调整, 主要可分为两个方面: 一是对模糊集、权重等参数进行优化, 确保规则库的明确语义、一致性、完备性及物理意义等; 二是对规则库进行简化, 获得合适规模的模型, 确保其紧凑性与可读性.

Jin 等在进化算法的适应度函数中加入了可分辨性、完备性和一致性度量, 提出了一种利用进化策略从数据中生成可区分的、完整的、一致的、紧凑的模糊规则 (Distinguishable, complete, consistent and compact fuzzy rule, DC^3 fuzzy rule) 的方法, 该研究表明, 当训练数据较少时, 提高训练后模型的可解释性有利于提高其综合性能[75]. 文献 [104] 在设计进化算法时, 将对可解释性的寻优转化为相应的约束条件, 提出了一种基于协同进化的模糊建模技术 (Fuzzy CoCo), 实现了对隶属函数和规则的两个独立但相互交织的搜索过程. 文献 [105] 考虑了所得规则库的透明性及工程所需的逼近精度, 提出了一种可靠的数据驱动方法来确定模糊系统的结构及参数, 例如输入变量及其隶属函数等. 在 BRB

建模领域, 通过自适应优化方法保证模型可解释性的相关研究还不够系统. Yang 等首次提出了 BRB 模型参数离线优化方法, 并强调专家知识是 BRB 可解释的基础, BRB 的参数寻优过程应该是一个基于专家初始判断的局部过程[106−107]. Zhou 等首次提出了 BRB 参数在线更新方法, 并结合实际系统的机理信息在参数更新过程中引入了专家干预, 有效保证了优化过程的可解释性[108]. 此外, Zhou 等还首次提出了 BRB 结构在线更新算法, 有效减小了规则库的规模[74]. 基于上述研究, Chang 等首次提出了 BRB 参数和结构联合优化模型, 为建立更加紧凑的规则库模型奠定了基础[109].

2.4 其他解释方法

上述研究致力于提高模型的 Ante-hoc 可解释性. 事实上, 其 Post-hoc 可解释性也受到较多的关注, 常见的方法有对模型的敏感性分析 (Sensitivity analysis) 与局部近似分析 (Local approximate analysis). 敏感性分析指在给定假设下定量地研究自变量发生某种变化对某一特定的因变量影响程度的一种不确定分析技术. Feng 等通过求导建立了输入与可靠度、权重等 BRB 参数的关系, 实现了对 BRB 模型参数的敏感性分析, 该方法能够为工程实际中的复杂系统维护决策提供指导[24]. Yang 等通过对 BRB 系统的匹配度及激活权值等参数进行了敏感性分析, 考察了效用值和属性权重对系统准确性的影响, 并提出了一种新的激活权值计算方法和参数优化方法, 提高了 BRB 系统的可解释性[25]. 文献 [110] 提出了模糊集最大扰动和平均扰动的定义, 并讨论了若干模糊推理方法的扰动结果. 文献 [111] 基于连接词和蕴涵算子对模糊推理方法进行了敏感性分析. 局部近似分析一般指分析模型局部的逼近过程, 然后基于此可对模型全局推理过程进行分析. Chen 等对 BRB 系统的推理和近似性质进行了理论分析, 分析结果揭示了 BRB 系统的推理机制拥有优越的逼近性能, 即 BRB 系统的统一多模型分解结构 (The unified multi-model decomposition structure) 将输入空间划分成不同的局部区域然后进行分布逼近[112]. Chen 等的研究为实际应用中 BRB 系统的使用和优化提供了理论基础.

3 RBM 可解释性的应用

建模方法的可解释性具有重要的理论和实践价值, 它有助于加深使用者对模型的理解和信任以开展进一步的应用活动, 例如故障诊断、缓解过拟合和模型迁移. 此外, 可解释性还有益于发现新知识, 甚至促成新理论.

总而言之, RBM 可解释性的工程意义具体可以体现在以下几个应用领域:

1) 辅助分析与决策. RBM 可解释性有利于提高人工分析和决策的效率, 提高分析与决策结果的可信度. 在医疗领域, RBM 一般会作为专家系统辅助医生做出正确决策. 2002 年, Yuan 等基于模糊逻辑推理方法开发了肾移植指派专家系统[113]. 2009 年, Kong 等构建了基于置信规则的临床决策支持系统 (Clinical decision support systems, CDSS), 实现了临床医疗的现场决策[114]. Zhou 等还提出了一种用于诊断胃癌淋巴结转移的决策支持系统[115]. Hossain 等针对患者体征和症状的不确定性, 基于规则库专家系统建立了急性冠脉综合征的评估模型, 该模型在后面的研究里还用于诊断结核病[116]. 在工商业领域, RBM 可以用来对生产过程提供指导和预测. 2012~2014 年间, Zalnezhad 等基于模糊规则实现了航空航天 AL7075-T6 合金 Ti-N 涂层结合强度预测[117]. 同时还建立了 AL7075-T6 合金钛锡镀层表面粗糙度随直流电源、温度、直流偏压、氮气流量等输入工艺参数变化的模糊模型, 实现了涂层试样的微动疲劳寿命的预测, 依靠所建立的模型有效改善了生产工艺, 提高了产品品质[118]. Chen 等开发了一个具有非线性现金流约束的投资组合优化的 BRB 系统, 实现了投资风险的辅助决策[119].

2) 复杂系统健康管理. RBM 的可解释性保证了所建立的模型具有输出的可追溯性, 即可依据输出对输入的相关情况进行分析, 这在复杂系统的健康管理中具有重要的作用. Xie 等基于模糊规则建立了预警机故障预测与健康管理系统, 不仅可以增强预警机健康状态监测能力, 而且可以提高其雷达故障诊断和维修的效率[120]. Ishibashi 等基于规则的遗传模糊系统 (Genetic fuzzy rule- based system, GFRBS) 混合模型提出了一种通用系统剩余使用寿命的预测方法, 该方法自动生成模糊规则, 并对关联的隶属度函数进行调优[121]. 该方法应用于商用航空飞机发动机的剩余使用寿命预测取得了较好的效果. 2015 年, Zhou 等提出了一种用于复杂系统行为预测的模型, 该模型最大的优势在于可以利用客观世界的专家经验, 使专家参与到操作过程中[122]. 2016 年, Hu 等提出了一个 BRB 预测模型来预测网络安全状况这一隐含行为, 以规则的形式展示了网络安全状况的演化过程[123]. Feng 等基于 BRB 系统构建了用于 WD615 型柴油发动机安全性评估的专家系统, 通过敏感性分析方法得到了影响发动机安全性的关键因素, 为下一步该型发动机的维护保

养提供支持[24]. Zhou 等基于前期研究, 开发了复杂系统安全性推演演示验证系统, 如图 5 所示. 该系统的关键是依据复杂系统的故障树、设计资料、专家经验以及测试数据构建规则库, 并采用 ER 推理方法生成结果. 同时, 用户可以依据输出结果和激活规则对输入信息进行分析, 从而确定复杂系统安全性薄弱环节.

图 5　复杂系统安全性推演演示验证系统

Fig. 5　Complex system security demonstration and verification system

3) 知识发现 (Knowledge discovery in database, KDD), 是指从数据中识别出有效的、新颖的、潜在有用的、最终可理解的模式的过程. RBM 在知识发现的优势主要为其具有较高的性能并且易于用户理解, 通常不需要进行其他可视化操作. 常见的应用情况是采用 RBM 从清洗后的海量数据中提取数据映射模式, 最后利用规则将挖掘到的数据模式可视化地呈现给用户. 基于规则的知识发现方法主要包括决策树、基于模糊规则的聚类方法和基于模糊规则的数据粒化方法[1].

4) 基于规则的控制. RBM 具有实时性好, 操作简单且计算量小等特点, 在工程中可以避免复杂控制系统难以建立精确数学模型的问题, 在控制领域应用广泛. 牛培峰等采用两层模糊规则控制方法实现了对循环流化床床温的控制系统[124]. 张海等人提出一种全新的基于模糊推理与规则控制的高度跟踪算法, 该算法具有较好的实时性, 收敛性及可扩展性, 在跟踪效果与运算速度上均优于传统算法, 对巡航导弹、无人机具有较高的应用价值[125].

此外, RBM 还可用于模型验证与诊断. 传统的验证方法通常基于模型在验证集上的误差来评估其泛化性能, 但当所用数据集存在偏差或验证集与训练集同分布时, 上述方法变得不可靠. RBM 的可解释性可作为一种可靠的依据来对模型进行分析和调试, 以诊断出模型中存在的缺陷并采取相应人为干预, 避免产生错误决策.

4　RBM 面临的问题及未来发展

RBM 由于其较好的可解释性与建模精度得到了广泛应用. 但在工程实践中, 仍旧存在几个问题:

1) 忽略了专家初始判断在建模过程中的重要性. 专家知识是基于规则建模方法的可解释性的重要来源, 它能辅助确定模型的结构、参数数量与大小. 专家知识准确与否、可信与否影响着所建模型的精确性与可解释性, 专家知识的局限性同样会制约所建模型的性能. 因此, RBM 需采用一系列优化算法进行寻优, 但在其过程中应充分考虑专家初始判断, 避免纯数据驱动的寻优过程.

2) 处理高维信息的能力较弱. 随着实际工程系统不断复杂化, 建模过程中对高维信息的处理越来越难以避免. RBM 在处理高维信息时面临的关键问题是 “维数灾难”, 尽管已有研究能较大程度上减小规则库的规模, 但是目前的研究仍不具有足够的一般性. 例如特征选择方法可以辅助剔除大量冗余特征, 但是这样的剔除过程是否合理, 是否适用于所有情况仍缺少足够的依据. “维数灾难” 带来的典型问题有: 如何获取海量规则信息? 如何解释复杂规则库? RBM 没有内置的学习机制, 如何对海量规则进行优化?

3) 可扩展能力较弱. 随着人们对实际工程系统认识的不断加深, 新的组成 (属性、参考值、规则等) 被期望引入原有规则库的同时应尽量避免对原有规则库的更改. 现有研究对参考值、规则等的增减进行了研究, 一定程度上保障了规则库扩展的合理性, 但是鲜有研究考虑到属性输入数量的增减. RBM 的适应性调整有利于建立更加完备可靠的知识库, 对用户理解和使用 RBM 具有较好的辅助作用. 因此, 如何建立更具一般扩展性的规则模型是一个需要解决的实际问题.

4) 缺乏统一的可解释性评估方法. RBM 可解释性研究领域缺乏统一的科学评估体系用于评估解释方法的优劣. 现有许多评估方法是依赖于人类认知的定性评估方法, 难以保证评估结果的可靠性. 此外, 迫切需要构建统一的可解释性评估方法, 以统一 RBM 的性能衡量, 这将为其在实践过程中的进一步应用提供帮助.

RBM 的优势在于其具有基于语义实现复杂非线性建模的可解释性. 从 RBM 自身而言, 未来围绕上述四个问题的发展方向都将对提高其建模能力具有较大的促进作用. 另一方面, 对 RBM 的扩展或者与其他方法的结合也是一个重要的发展方向.

Zhou 等突破深度学习依赖于神经网络的困境, 提出了一种全新的深度学习框架—深度随机森林模型 (Deep random forest)[126]. 该模型由大量决策树组成, 决策树模型可以视作基于特殊二分规则的模型. 与神经网络模型相比, 决策树的二分规则更易于解释, 且更容易形成清晰的推理过程[35−36]. 深度随机森林模型为以 RBM 为基础进一步探索可解释的深度学习框架提供了指导. 除此之外, 受模糊神经网络启发, 也有研究尝试将规则嵌入神经网络, 从一定程度上提高了模型的透明性, 增强了模型的局部解释能力. 在未来的研究里, 如何进一步融合 RBM 与其他建模方法并推动其向深度发展, 仍旧是一个很有价值的研究课题.

5 结论

建模方法的可解释性在理论与工程领域具有很高的研究价值, 已经成为了国内外学者研究的热点问题, 并且取得了较好的研究成果. RBM 作为一种可自解释的建模方法, 在可解释性研究领域受到广泛的关注. 本文总结分析了 RBM 的发展过程, 从知识库、推理机、自适应学习方法以及其他解释方法等方面系统梳理了国内外与 RBM 可解释性相关的典型工作, 同时对 RBM 的工程应用进行了简要的总结, 为其发展和改进提供了一定的参考和借鉴. 从本文的总结可以看出, RBM 的可解释性研究还处于初级阶段, 依然存在许多关键问题尚待解决, 例如处理高维信息的能力弱, 缺乏统一的可解释性评估方法等. 针对这些问题, 本文所讨论的未来的研究方向, 有利于进一步推动 RBM 可解释性研究, 提高其建模性能.

References

1 Casillas J, Cordón O, Herrera F, Magdalena L. *Interpretability Issues in Fuzzy Modeling*. Berlin, Heidelberg: Springer, 2003.

2 He R, Wu X, Sun Z N, Tan T N. Wasserstein CNN: Learning invariant features for NIR-VIS face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019, **41**(7): 1761−1773

3 Miao K H, Miao J H. Coronary heart disease diagnosis using deep neural networks. *International Journal of Advanced Computer Science and Applications*, 2018, **9**(10): 1−8

4 Ogunfunmi T, Ramachandran R P, Togneri R, Zhao Y J, Xia X J. A primer on deep learning architectures and applications in speech processing. *Circuits, Systems, and Signal Processing*, 2019, **38**(8): 3406−3432

5 Hori T, Chen Z, Erdogan H, Hershey J R, Le Roux J, Mitra V, et al. Multi-microphone speech recognition integrating beamforming, robust feature extraction, and advanced DNN/RNN backend. *Computer Speech and Language*, 2017, **46**: 401−418

6 Wyatt J. Nervous about artificial neural networks. *The Lancet*, 1995, **346**(8984): 1175−1177

7 Walker C R, Frize M. Are Artificial Neural Networks "Ready to Use" for Decision Making in the Neonatal Intensive Care Unit?: Commentary on the article by Mueller et al. and page 11. *Pediatric Research*, 2004, **56**(1): 6−8

8 Gacto M J, Alcalá R, Herrera F. Interpretability of linguistic fuzzy rule-based systems: An overview of interpretability measures. *Information Sciences*, 2011, **181**(20): 4340−4360

9 Guillaume S. Designing fuzzy inference systems from data: An interpretability-oriented review. *IEEE Transactions on Fuzzy Systems*, 2001, **9**(3): 426−443

10 Zhou S M, Gan J Q. Low-level interpretability and high-level interpretability: A unified view of data-driven interpretable fuzzy system modelling. *Fuzzy Sets and Systems*, 2008, **159**(23): 3091−3131

11 Mencar C, Fanelli A M. Interpretability constraints for fuzzy information granulation. *Information Sciences*, 2008, **178**(24): 4585−4618

12 Jin Y. Fuzzy modeling of high-dimensional systems: Complexity reduction and interpretability improvement. *IEEE Transactions on Fuzzy Systems*, 2000, **8**(2): 212−220

13 Montavon G, Samek W, Müller K R. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 2018, **73**(8): 1−15

14 Robinson R L, Palczewska A, Palczewski J, Kidley N. Comparison of the predictive performance and interpretability of random forest and linear models on benchmark data sets. *Journal of Chemical Information and Modeling*, 2017, **57**(8): 1773−1792

15 Zhang Y R, Haghani A. A gradient boosting method to improve travel time prediction. *Transportation Research Part C: Emerging Technologies*, 2015, **58**: 308−324

16 Cireşan D, Meier U, Masci J, Schmidhuber J. Multi-column deep neural network for traffic sign classification. *Neural Networks*, 2012, **32**: 333−338

17 Andrews R, Diederich J, Tickle A B. Survey and critique of techniques for extracting rules from trained artificial neural networks. *Knowledge-Based Systems*, 1995, **8**(6): 373−389

18 Tickle A B, Andrews R, Golea M, Diederich J. The truth will come to light: Directions and challenges in extracting the knowledge embedded within trained artificial neural networks. *IEEE Transactions on Neural Networks*, 1998, **9**(6): 1057−1068

19 Fu L M. Rule generation from neural networks. *IEEE Transactions on Systems, Man, and Cybernetics*, 1994, **24**(8): 1114−1124

20 De Fortuny E J, Martens D. Active learning-based pedagogical rule extraction. *IEEE Transactions on Neural Networks and Learning Systems*, 2015, **26**(11): 2664−2677

21 Wang J P, Gou L, Zhang W, Yang H, Shen H W. *DeepVID*: Deep visual interpretation and diagnosis for image classifiers via knowledge distillation. *IEEE Transactions on Visualization and Computer Graphics*, 2019, **25**(6): 2168−2180

22 Chen Y, Meng H B, Wen X L, Ma P G, Qin Y X, Ma Z X, et al. Classification methods of a small sample target object in the sky based on the higher layer visualizing feature and transfer learning deep networks. *EURASIP Journal on Wireless Communications and Networking*, 2018, 2018: Article No. 127

23 Oberguggenberger M, King J, Schmelzer B. Classical and imprecise probability methods for sensitivity analysis in engineering: A case study. *International Journal of Approximate Reasoning*, 2009, **50**(4): 680−693

24 Feng Z C, Zhou Z J, Hu C H, Chang L L, Hu G Y, Zhao F J. A new belief rule base model with attribute reliability. *IEEE Transactions on Fuzzy Systems*, 2019, **27**(5): 903−916

25 Yang L H, Liu J, Wang Y M, Martínez L. New activation weight calculation and parameter optimization for extended belief rule-based system based on sensitivity analysis. *Knowledge and Information Systems*, 2019, **60**(2): 837−878

26 Jensen C A, Reed R D, Marks R J, El-Sharkawi M A, Jung J B, Miyamoto R T, et al. Inversion of feedforward neural networks: Algorithms and applications. *Proceedings of the IEEE*, 1999, **87**(9): 1536−1549

27 Lee G, Jeong J, Seo S, Kim C, Kang P. Sentiment classification with word localization based on weakly supervised learning with a convolutional neural network. *Knowledge-Based Systems*, 2018, **152**(15): 70−82

28 Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D. A survey of methods for explaining black box models. *ACM Computing Surveys*, 2018, **51**(5): Article No. 93

29 Guo H P, Wu S H, Wang Z Q, Wu C A. Linear regression for forecasting photovoltaic power generation. *Applied Mechanics and Materials*, 2014, **494-495**: 1771−1774

30 Sun H M. A naive bayes classifier for prediction of multidrug resistance reversal activity on the basis of atom typing. *Journal of Medicinal Chemistry*, 2005, **48**(12): 4031−4039

31 Afsari F, Eftekhari M, Eslami E, Woo P Y. Interpretability-based fuzzy decision tree classifier a hybrid of the subtractive clustering and the multi-objective evolutionary algorithm. *Soft Computing*, 2013, **17**(9): 1673−1686

32 Sun R. Robust reasoning: Integrating rule-based and similarity-based reasoning. *Artificial Intelligence*, 1995, **75**(2): 241−295

33 Haufe S, Meinecke F, Görgen K, Dähne S, Haynes J D, Blankertz B, et al. On the interpretation of weight vectors of linear models in multivariate neuroimaging. *NeuroImage*, 2014, **87**: 96−110

34 Štrumbelj E, Kononenko I. An efficient explanation of individual classifications using game theory. *Journal of Machine Learning Research*, 2010, **11**(1): 1−18

35 Huysmans J, Dejaeger K, Mues C, Vanthienen J, Baesens B. An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models. *Decision Support Systems*, 2011, **51**(1): 141−154

36 Breslow L A, Aha D W. Simplifying decision trees: A survey. *The Knowledge Engineering Review*, 1997, **12**(1): 1−40

37 Breiman L. Random forests. *Machine Learning*, 2001, **45**(1): 5−32

38 Zadeh L A. Fuzzy sets. *Information and Control*, 1965, **8**(3): 338−353

39 Mamdani E H. Application of fuzzy algorithms for control of simple dynamic plant. *Proceedings of the Institution of Electrical Engineers*, 1974, **121**(12): 1585−1588

40 Fiordaliso A. A constrained Takagi-Sugeno fuzzy system that allows for better interpretation and analysis. *Fuzzy Sets and Systems*, 2001, **118**(2): 307−318

41 Bikdash M. A highly interpretable form of Sugeno inference systems. *IEEE Transactions on Fuzzy Systems*, 1999, **7**(6): 686−696

42 Zhou Zhi-Jie, Chen Yu-Wang, Hu Chang-Hua, Zhang Bang-Cheng, Chang Lei-Lei. *Evidential Reasoning, Belief Rule Base and Complex System Modeling*. Beijing: Science Press, 2017.
(周志杰, 陈玉旺, 胡昌华, 张邦成, 常雷雷. 证据推理、置信规则库与复杂系统建模. 北京: 科学出版社, 2017.)

43 Yang J B, Sen P. A general multi-level evaluation process for hybrid MADM with uncertainty. *IEEE Transactions on Systems, Man, and Cybernetics*, 1994, **24**(10): 1458−1473

44 Yang J B, Singh M G. An evidential reasoning approach for multiple-attribute decision making with uncertainty. *IEEE Transactions on Systems, Man, and Cybernetics*, 1994, **24**(1): 1−18

45 Yang J B, Liu J, Wang J, Sii H S, Wang H W. Belief rule-base inference methodology using the evidential reasoning approach-RIMER. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 2006, **36**(2): 266−285

46 Dempster A P. A generalization of Bayesian inference. *Journal of the Royal Statistical Society. Series B: Methodological*, 1968, **30**(2): 205−247

47 Shafer G. *A Mathematical Theory of Evidence*. Princeton: Princeton University Press, 1976.

48 Poulton E C. *Behavioral Decision Theory*. Cambridge: Cambridge University Press, 1994.

49 Goodman I R, Nguyen HT. Uncertainty Models for Knowledge Based Systems. *North HollandPublishing Co Amsterdam*, 1991

50 Liu J, Yang J B, Wang J, Sii H S, Wang Y M. Fuzzy rule-based evidential reasoning approach for safety analysis. *International Journal of General Systems*, 2004, **33**(2-3): 183−204

51 Post E L. Introduction to a general theory of elementary propositions. *American Journal of Mathematics*, 1921, **43**(3): 163−185

52 Wang Ning, Meng Xian-Yao. Stability analysis of T-S fuzzy control system with inputs using general fuzzy partition. *Acta Automatica Sinica*, 2008, **34**(11): 1441−1445
(王宁, 孟宪尧. 输入采用一般模糊划分的 T-S 模糊控制系统稳定性分析. 自动化学报, 2008, **34**(11): 1441−1445)

53 Zhang Song-Tao, Ren Guang. Analysis and design of discrete fuzzy system based on piecewise fuzzy Lyapunov approach. *Acta Automatica Sinica*, 2006, **32**(5): 813−818
(张松涛, 任光. 基于分段模糊 Lyapunov 方法的离散模糊系统分析与设计. 自动化学报, 2006, **32**(5): 813−818)

54 Chang L L, Zhou Z J, You Y, Yang L H, Zhou Z G. Belief rule based expert system for classification problems with new rule activation and weight calculation procedures. *Information Sciences*, 2016, **336**: 75−91

55 Chang L L, Zhou Z J, Liao H C, Chen Y W, Tan X, Herrera F. Generic disjunctive belief-rule-base modeling, inferencing, and optimization. *IEEE Transactions on Fuzzy Systems*, 2019, **27**(9): 1866−1880

56 Chang L L, Jiang J, Sun J B, Chen Y W, Zhou Z J, Xu X B, et al. Disjunctive belief rule base spreading for threat level assessment with heterogeneous, insufficient, and missing information. *Information Sciences*, 2019, **476**: 106−131

57 Chang L L, Chen Y W, Hao Z Y, Zhou Z J, Xu X B, Tan X. Indirect disjunctive belief rule base modeling using limited conjunctive rules: Two possible means. *International Journal of Approximate Reasoning*, 2019, **108**: 1−20

58 Liu J, Martinez L, Wang Y M. Extended belief rule base inference methodology. In: Proceedings of the 3rd International Conference on Intelligent System and Knowledge Engineering. Xiamen, China: IEEE, 2008. 1415−1420

59 Liu J, Martinez L, Calzada A, Wang H. A novel belief rule base representation, generation and its inference methodology. *Knowledge-Based Systems*, 2013, **53**: 129−141

60 Liao Gui-Min. A survey of knowledge representation methods based on fault tree model. *Computer and Information Technology*, 2000, (1): 6−8, 52
(廖贵敏. 基于故障树模型的知识表达方法综述. 电脑与信息技术, 2000, (1): 6−8, 52)

61 Tang D W, Yang J B, Chin K S, Wong Z S Y, Liu X B. A methodology to generate a belief rule base for customer perception risk analysis in new product development. *Expert Systems with Applications*, 2011, **38**(5): 5373−5383

62 Riid A. Transparent Fuzzy Systems: Modeling and Control [Ph. D. dissertation], Tallinn Technical University, Estonia, 2002

63 De Oliveira J V. A design methodology for fuzzy system interfaces. *IEEE Transactions on Fuzzy Systems*, 1995, **3**(4): 404−414

64 De Oliveira J V. Semantic constraints for membership function optimization. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 1999, **29**(1): 128−138

65 Dubois D, Prade H. *Fuzzy Sets and Systems: Theory and Ap-*

plications. New York: Academic Press, 1980.

66 Zwick R, Carlstein E, Budescu D V. Measures of similarity among fuzzy concepts: A comparative analysis. *International Journal of Approximate Reasoning*, 1987, **1**(2): 221–242

67 Mencar C, Castellano G, Fanelli A M. Distinguishability quantification of fuzzy sets. *Information Sciences*, 2007, **177**(1): 130–149

68 Mencar C, Castellano G, Bargiela A, Fanelli A M. Similarity vs. possibility in measuring fuzzy sets distinguishability. In: Proceedings of the 5th International Conference on Recent Advances in Soft Computing. Nottingham, UK: Nottingham Trent University, 2004. 354–359

69 Zhou S M, Gan J Q. Constructing accurate and parsimonious fuzzy models with distinguishable fuzzy sets based on an entropy measure. *Fuzzy Sets and Systems*, 2006, **157**(8): 1057–1074

70 Casillas J, Martínez P, Benítez A D. Learning consistent, complete and compact sets of fuzzy rules in conjunctive normal form for regression problems. Soft Computing, 2009, **13**(5): 451–465

71 Meesad P, Yen G G. Accuracy, comprehensibility and completeness evaluation of a fuzzy expert system. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 2003, **11**(4): 445–466

72 Espinosa J, Vandewalle J. Constructing fuzzy models with linguistic integrity from numerical data-AFRELI algorithm. *IEEE Transactions on Fuzzy Systems*, 2000, **8**(5): 591–600

73 Li G L, Zhou Z J, Hu C H, Chang L L, Zhou Z G, Zhao F J. A new safety assessment model for complex system based on the conditional generalized minimum variance and the belief rule base. *Safety Science*, 2017, **93**: 108–120

74 Zhou Z J, Hu C H, Yang J B, Xu D L, Chen M Y, Zhou D H. A sequential learning algorithm for online constructing belief-rule-based systems. *Expert Systems With Applications*, 2010, **37**(2): 1790–1799

75 Jin Y C, von Seelen W, Sendhoff B. On generating FC3 fuzzy rule systems from data using evolution strategies. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 1999, **29**(6): 829–845

76 Ding C, Peng H C. Minimum redundancy feature selection from microarray gene expression data. *Journal of Bioinformatics and Computational Biology*, 2005, **3**(2): 185–205

77 Chang L L, Zhou Y, Jiang J, Li M J, Zhang X H. Structure learning for belief rule base expert system: A comparative study. *Knowledge-Based Systems*, 2013, **39**: 159–172

78 Chang Lei-Lei, Li Meng-Jun, Lu Yan-Jing, Cheng Ben, Zhang Xiao-Hang. Structure learning for belief rule base using principal component analysis. *Systems Engineering: Theory and Practice*, 2014, **34**(5): 1297–1304
(常雷雷, 李孟军, 鲁延京, 程贲, 张晓航. 基于主成分分析的置信规则库结构学习方法. 系统工程理论与实践, 2014, **34**(5): 1297–1304)

79 Wang Ying-Ming, Yang Long-Hao, Chang Lei-Lei, Fu Yang-Geng. Rough set method for rule reduction in belief rule base. *Control and Decision*, 2014, **29**(11): 1943–1950
(王应明, 杨隆浩, 常雷雷, 傅仰耿. 置信规则库规则约简的粗糙集方法. 控制与决策, 2014, **29**(11): 1943–1950)

80 Chang L L, Zhou Z J, Chen Y W, Xu X B, Sun J B, Liao T J, et al. Akaike Information Criterion-based conjunctive belief rule base learning for complex system modeling. *Knowledge-Based Systems*, 2018, **161**: 47–64

81 Jin Y C. Generating distinguishable, complete, consistent and compact fuzzy systems using evolutionary algorithms. *Accuracy Improvements in Linguistic Fuzzy Modeling. Studies in Fuzziness and Soft Computing*. Berlin, Heidelberg: Springer, 2003. 100–118

82 Hodges J, Bridges S, Sparrow C, Wooley B, Tang B, Jun C. The development of an expert system for the characterization of containers of contaminated waste. *Expert Systems with Applications*, 1999, **17**(3): 167–181

83 Walley P. Measures of uncertainty in expert systems. *Artificial Intelligence*, 1996, **83**(1): 1–58

84 Walley P. *Statistical Reasoning with Imprecise Probabilities*. New York: Chapman and Hall, 1991.

85 Burnside W. *Theory of Probability*. Cambridge: Cambridge University Press, 1928.

86 Chapman V. Making decisions. *Nurs Stand*, 1995, **10**(8): 2–8

87 Andersen S K. Probabilistic reasoning in intelligent systems: Networks of plausible inference: Judea Pearl. *Artificial Intelligence*, 1991, **48**(1): 117–124

88 Nilsson N J. Probabilistic logic. *Artificial Intelligence*, 1986, **28**(1): 71–87

89 Duda R O, Reboh R. AI and decision makings: The prospector experience. *Artificial Intelligence Applications for Business*, 1984, **21**: 111–147

90 Lauritzen S L, Spiegelhalter D J. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society. Series B: Methodological*, 1998, **50**(2): 157–224

91 Spiegelhalter D J, Dawid A P, Lauritzen S L, Cowell R G. Bayesian analysis in expert systems. *Statistical Science*, 1993, **8**(3): 219–247

92 Pearl J. Reasoning with belief functions: An analysis of compatibility. *International Journal of Approximate Reasoning*, 1990, **4**(5-6): 363–389

93 Halpern J Y, Fagin R. Two views of belief: Belief as generalized probability and belief as evidence. *Artificial Intelligence*, 1992, **54**(3): 275–317

94 Shafer G. Constructive probability. *Synthese*, 1981, **48**(1): 1–60

95 Shafer G. Belief functions and parametric models. *Journal of the Royal Statistical Society. Series B: Methodological*, 1982, **44**(3): 322–339

96 Shafer G. Perspectives on the theory and practice of belief functions. *International Journal of Approximate Reasoning*, 1990, **4**(5-6): 323–362

97 Voorbraak F. A computationally efficient approximation of Dempster-Shafer theory. *International Journal of Man-machine Studies*, 1989, **30**(5): 525–536

98 Tzvieli A. Possibility theory: An approach to computerized processing of uncertainty. *Journal of the American Society for Information Science*, 1990, **41**(2): 153–154

99 Yager R R. An introduction to applications of possibility theory (+ commentaries by L.A. Zadeh, W. Bandler, T. Saaty, A. Kandel, D. Dubois & H. Prade, R.M. Tong and M. Kochen). *Human Systems Management*, 1982, **3**(4): 246–269

100 Dubois D, Prade H. *Possibility Theory*. New York: Plenum Press, 1988.

101 Smith C A B. Consistency in statistical inference and decision. *Journal of the Royal Statistical Society. Series B: Methodological*, 1961, **23**(1): 1–25

102 Jaffray J Y. Bayesian updating and belief functions. *IEEE Transactions on Systems, Man, and Cybernetics*, 1992, **22**(5): 1144–1152

103 Wang Y M, Yang J B, Xu D L. Environmental impact assess-

ment using the evidential reasoning approach. *European Journal of Operational Research*, 2006, **174**(3): 1885−1913

104 Pena-Reyes C A, Sipper M. Fuzzy CoCo: A cooperative-coevolutionary approach to fuzzy modeling. *IEEE Transactions on Fuzzy Systems*, 2001, **9**(5): 727−737

105 Guillaume S, Charnomordic B. Generating an interpretable family of fuzzy partitions from data. *IEEE Transactions on Fuzzy Systems*, 2004, **12**(3): 324−335

106 Yang J B, Liu J, Xu D L, Wang J, Wang H W. Optimization models for training belief-rule-based systems. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 2007, **37**(4): 569−585

107 Xu D L, Liu J, Yang J B, Liu G P, Wang J, Jenkinson I, et al. Inference and learning methodology of belief-rule-based expert system for pipeline leak detection. *Expert Systems with Applications*, 2007, **32**(1): 103−113

108 Zhou Z J, Hu C H, Yang J B, Xu D L, Zhou D H. Online updating belief-rule-base using the RIMER approach. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 2011, **41**(6): 1225−1243

109 Chang L L, Zhou Z J, Chen Y W, Liao T J, Hu Y, Yang L H. Belief rule base structure and parameter joint optimization under disjunctive assumption for nonlinear complex system modeling. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2018, **48**(9): 1542−1554

110 Ying M S. Perturbation of fuzzy reasoning. *IEEE Transactions on Fuzzy Systems*, 1999, **7**(5): 625−629

111 Cai K Y. Robustness of fuzzy reasoning and δ-equalities of fuzzy sets. *IEEE Transactions on Fuzzy Systems*, 2001, **9**(5): 738−750

112 Chen Y W, Yang J B, Xu D L, Yang S L. On the inference and approximation properties of belief rule based systems. *Information Sciences*, 2013, **234**: 121−135

113 Yuan Y F, Feldhamer S, Gafni A, Fyfe F, Ludwin D. The development and evaluation of a fuzzy logic expert system for renal transplantation assignment: Is this a useful tool? *European Journal of Operational Research*, 2002, **142**(1): 152−173

114 Kong G L, Xu D L, Liu X B, Yang J B. Applying a belief rule-base inference methodology to a guideline-based clinical decision support system. *Expert Systems*, 2009, **26**(5): 391−408

115 Zhou Z G, Liu F, Jiao L C, Zhou Z J, Yang J B, Gong M G, et al. A bi-level belief rule based decision support system for diagnosis of lymph node metastasis in gastric cancer. *Knowledge-Based Systems*, 2013, **54**: 128−136

116 Hossain M S, Rahaman S, Mustafa R, Andersson K. A belief rule-based expert system to assess suspicion of acute coronary syndrome (ACS) under uncertainty. *Soft Computing*, 2018, **22**: 7571−7586

117 Zalnezhad E, Sarhan A A D M, Hamdi M. Prediction of TiN coating adhesion strength on aerospace AL7075-T6 alloy using fuzzy rule based system. *International Journal of Precision Engineering and Manufacturing*, 2012, **13**(8): 1453−1459

118 Zalnezhad E, Sarhan A A D. A fuzzy logic predictive model for better surface roughness of Ti-TiN coating on AL7075-T6 alloy for longer fretting fatigue life. *Measurement*, 2014, **49**: 256−265

119 Chen Y W, Poon S H, Yang J B, Xu D L, Zhang D X, Acomb S. Belief rule-based system for portfolio optimisation with nonlinear cash-flows and constraints. *European Journal of Operational Research*, 2012, **223**(3): 775−784

120 Xie G J, Yan S Q, Tang Z Y, Rui L. A PHM system for AEW radar based on AOPS-LSSVM prognostic algorithm and expert knowledge database. In: Proceedings of the 2010 Prognostics and System Health Management Conference. Macao, China: IEEE, 2010. 1−6

121 Ishibashi R, Júnior C L N. GFRBS-PHM: A Genetic Fuzzy Rule-Based System for PHM with improved interpretability. In: Proceedings of the 2013 IEEE Conference on Prognostics and Health Management (PHM). Gaithersburg, MD, USA: IEEE, 2013. 1−7

122 Zhou Z J, Hu C H, Hu G Y, Han X X, Zhang B C, Chen Y W. Hidden behavior prediction of complex systems under testing influence based on semiquantitative information and belief rule base. *IEEE Transactions on Fuzzy Systems*, 2015, **23**(6): 2371−2386

123 Hu G Y, Zhou Z J, Zhang B C, Yin X J, Gao Z, Zhou Z G. A method for predicting the network security situation based on hidden BRB model and revised CMA-ES algorithm. *Applied Soft Computing*, 2016, **48**: 404−418

124 Niu Pei-Feng, Ding Xi-Sheng. Application of double-deck fuzzy control to bed temperature control system of circulated fluidized-bed. *Journal of Yanshan University*, 2008, **32**(2): 124−128
(牛培峰, 丁希生. 两层模糊控制在循环流化床床温控制系统中的应用. 燕山大学学报, 2008, **32**(2): 124−128)

125 Zhang Hai, Zhou De-Yun, Tong Ming-An. A quick altitude following algorithm based on rules control. *Fire Control and Command Control*, 1999, **24**(3): 21−26
(张海, 周德云, 佟明安. 基于规则控制的快速高度跟踪算法. 火力与指挥控制, 1999, **24**(3): 21−26)

126 Zhou Z H, Feng J. Deep forest: Towards an alternative to deep neural networks. In: Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI-17). Melbourne, Australia, 2017. 3553−3559

周志杰 火箭军工程大学教授. 2010 年获得清华大学博士学位. 主要研究方向为证据推理, 置信规则库, 故障诊断, 安全性评估.
E-mail: zhouzj04@tsinghua.org.cn
(**ZHOU Zhi-Jie** Professor at Rocket Force University of Engineering. He received his Ph. D. degree from Tsinghua University in 2010. His research interest covers evidential reasoning, belief rule base, fault prognosis, safety assessment.)

曹 友 火箭军工程大学博士研究生. 2017 年获得哈尔滨理工大学学士学位. 主要研究方向为证据推理, 置信规则库, 安全性评估. 本文通信作者.
E-mail: cy936756268@163.com
(**CAO You** Ph. D. candidate at Rocket Force University of Engineering. He received his bachelor degree from Harbin University of Science and Technology in 2017. His research interest covers evidential reasoning, belief rule base, safety assessment. Corresponding author of this paper.)

胡昌华 火箭军工程大学教授, 长江学者. 1996 年获得西北工业大学博士学位. 主要研究方向为故障诊断, 寿命预测.
E-mail: hch66603@163.com
(**HU Chang-Hua** Professor at Rocket Force University of Engineering, Cheung Kong Scholar. He received his Ph. D. degree from Northwestern Polytechnical University in 1996. His research interest covers fault prognosis and life prediction.)

唐帅文 火箭军工程大学博士研究生. 2017 年获得火箭军工程大学学士学位. 主要研究方向为证据推理, 故障诊断, 安全性评估.
E-mail: tsw631845201@163.com
(**TAGN Shuai-Wen** Ph. D. candidate at Rocket Force University of Engineering. He received his bachelor degree from Rocket Force University of Engineering in 2017. His research interest covers evidential reasoning, fault prognosis, and safety assessment.)

张春潮 火箭军工程大学博士研究生. 2019 年获得长春理工大学学士学位. 主要研究方向为置信规则库, 故障诊断.
E-mail: zhang1875349@163.com
(**ZHANG Chun-Chao** Ph. D. candidate at Rocket Force University of Engineering. He received his bachelor degree from Changchun University of Since and Technology in 2019. His research interest covers belief rule base and fault prognosis.)

王 杰 火箭军工程大学博士研究生. 2018 年获得合肥工业大学学士学位. 主要研究方向为证据推理, 故障诊断, 安全性评估.
E-mail: wj2802877478@163.com
(**WANG Jie** Ph. D. candidate at Rocket Force University of Engineering. He received his bachelor degree from Hefei University of Technology in 2018. His research interest covers evidential reasoning, fault prognosis, and safety assessment.)