Journal of Computer Applications
计算机应用,2014,34(8):2155 -2160,2169
ISSN 1001-9081
CODEN JYIIDU
2014-08-10
http://www.joca.cn

文章编号:1001-9081(2014)08-2155-06
doi:10.11772/j.issn.1001-9081.2014.08.2155

数据驱动的置信规则库构建与推理方法

余瑞银,杨隆浩,傅仰耿*
(福州大学 数学与计算机科学学院,福州 350116)
(*通信作者电子邮箱 ygfu@qq.com)

摘　要:针对 Liu 等(LIU J, MARTINEZ L, CALZADA A, *et al*. A novel belief rule base representation, generation and its inference methodology. Knowledge-Based Systems, 2013, 53: 129 -141)提出的扩展置信规则库(BRB)推理精度不够高的问题,提出了一种改进的规则库构建与推理方法。在 Liu 等提出的规则库构建方法的基础上,给出了一种新的生成规则前件与计算规则权重的方法;同时为了避免大量不必要的规则被激活,引入 80/20 法则改进规则激活策略,并最终形成完整的置信规则库构建与推理方法。通过输油管道检漏的实例对所提方法的准确性和效率进行对比分析。实验结果表明,所提方法能够在保证低耗时的同时,将系统平均绝对误差(MAE)降低到 0.173 42,具有较高的效率和精度。

关键词:置信规则库;证据推理;规则库构建;80/20 法则;输油管道检漏
中图分类号: TP18; TP273.5　　**文献标志码**:A

Data driven construction and inference methodology of belief rule-base

YU Ruiyin, YANG Longhao, FU Yanggeng*
(*College of Mathematics and Computer Science, Fuzhou University, Fuzhou Fujian* 350116, *China*)

Abstract: Considering the problem of the low inference accuracy of the extended Belief Rule Base (BRB) which was proposed by Liu, etc (LIU J, MARTINEZ L, CALZADA A, *et al*. A novel belief rule base representation, generation and its inference methodology. Knowledge-Based Systems, 2013, 53: 129 -141), an improved method of rule-base construction and inference was proposed. This approach was based on the method of Liu's rule-base construction, and a new generation method of rule antecedents and a new calculation method of rule weights were provided. Subsequently, in order to avoid activating so many unnecessary rules, the 80/20 rule was introduced to improve the strategy of rule activation. Then an integrated construction and inference methodology of belief rule-base was formed. Finally, in order to validate the accuracy and efficiency of the new approach, the case study in pipeline leak detection was provided. The experimental results show that the proposed approach not only can keep lower time-consumption, but also can make the Mean Absolute Error (MAE) of system be reduced to 0.173 42. This proves that the new approach has high accuracy and efficiency.

Key words: Belief Rule Base (BRB); evidential reasoning; construction of rule base; 80/20 rule; pipeline leak detection

0　引言

现实世界中总存在着由模糊、不完整、不精确等引起的各种不确定性[1],为表示和处理这些不确定性,Yang 等[2]在 Dempster-Shafer 证据理论、决策理论和规则库的基础上提出了基于证据推理的置信规则库推理方法(Belief Rule-Base Inference Methodology using the Evidential Reasoning approach, RIMER)。作为一种对传统 IF-THEN 规则库系统的扩展,置信规则库(Belief Rule Base, BRB)系统已成功应用于多个领域,例如输油管道检漏、消费者偏好预测、库存控制、临床风险评估和投资组合优化等[3-7]。应用 RIMER 方法需事先构建初始 BRB,同时给出 BRB 各个参数的初始值,包括规则权重、前提属性权重、前提属性候选值及结果属性置信度等。然而,由于专家知识的局限性,初始 BRB 未能具备良好的模拟实际系统的能力,因此,Yang 等[8]提出了训练 BRB 参数的最优化模型,经过训练的 BRB,其推理结果的精度得到明显提高。

BRB 参数优化的基本思想是在满足各个参数约束条件的前提下,最小化 BRB 输出与实际输出之间的差值。其中,最简便的参数训练方法是使用 Matlab 的 FMINCON 函数,但使用 FMINCON 函数的训练过程十分耗时,并且收敛精度也不够理想。因此,常瑞等[9]设计了一种基于梯度法和二分法的 BRB 训练方法,取得了更好的训练结果;Zhou 等[10]提出了基于克隆选择算法的 BRB 参数训练方法,具有较好的全局搜索能力。

然而,上述方法大多都由专家根据经验事先给定初始 BRB,同时需要耗费大量的时间对其进行参数训练。因此,

收稿日期:2014-04-08;**修回日期**:2014-05-08。
基金项目:国家自然科学基金青年项目(61300026,61300104);国家杰出青年科学基金资助项目(70925004);国家自然科学基金面上项目(71371053);福建省教育厅 A 类科技项目(JA13036);福州大学科技发展基金资助项目(2014-XQ-26)。
作者简介:余瑞银(1990 -),男,福建福州人,硕士研究生,主要研究方向:智能决策、置信规则库推理; 杨隆浩(1990 -),男,福建南平人,硕士研究生,主要研究方向:智能决策、置信规则库推理; 傅仰耿(1981 -),男,福建泉州人,讲师,博士,CCF 会员,主要研究方向:不确定多准则决策、置信规则库推理、移动互联网。

Liu 等[11]130-135提出了一种扩展的置信规则库(Extended Belief Rule Base, EBRB),与传统 BRB 的主要区别在于规则前件采用了分布式的表示类型。同时,在此基础上,文献[11]提出了一种由样本数据直接生成 EBRB 的方法,该方法无需专家给定初始规则库,也无需耗时的参数训练迭代过程,在很大程度上提高了推理的速度。然而,不合理的规则激活方法致使该方法推理精度不高,受噪声数据影响明显。鉴于此,本文在 Liu 等[11]方法的基础上,提出数据驱动的 BRB 构建与推理方法,首先将改进的由样本数据生成规则库的方法应用于传统的 BRB 上;然后在判断激活规则时,引入 80/20 法则[12]筛选相对重要的激活规则,提高 BRB 系统的推理精度;最后在输油管道检漏实验中对本文方法进行效率和误差分析,并与 Liu 等[11]和 Chen 等[13]方法进行对比,验证本文方法有效性和合理性。

1 BRB 和 EBRB

1.1 传统的 BRB 结构表示

在传统的产生式规则的基础上,Yang 等[2]269通过引入分布式置信框架和权重参数提出了置信规则的表示形式,其具体表示如下:

$$R_k:\text{IF } A_1^k \wedge A_2^k \wedge \cdots \wedge A_{T_k}^k \text{ THEN } \{(D_1,\bar{\beta}_{1k}),(D_2,\bar{\beta}_{2k}),\cdots,(D_N,\bar{\beta}_{Nk})\};\quad \sum_{i=1}^{N}\bar{\beta}_{ik}\leqslant 1 \tag{1}$$

其中:规则权重为 θ_k;属性权重为 $\delta_{k1},\delta_{k2},\cdots,\delta_{kT_k}(k\in\{1,2,\cdots,L\})$;$R_k$ 代表第 k 条规则;$A_i^k(i=1,2,\cdots,T_k,k=1,2,\cdots,L)$ 代表第 k 条规则的第 i 个前提属性(假设为 U_i)的候选值,T_k 代表第 k 条规则的前提属性个数,L 代表规则的总条数;$D_i(i=1,2,\cdots,N)$ 代表结果属性的第 i 个评价等级,N 代表评价等级的个数;$\bar{\beta}_{ik}(i=1,2,\cdots,N,k=1,2,\cdots,L)$ 代表在第 k 条规则中相对于评价等级 D_i 的置信度。若 $\sum_{i=1}^{N}\bar{\beta}_{ik}=1$ 时,表明第 k 条规则包含的信息是完整的;否则,表明该规则包含的信息是不完整的。

对于传统的 BRB 参数的取值,其初值通常是由专家给定,并通过对 BRB 参数学习方法进行优化,其中参数学习时所包含的参数有 θ_k、A_i^k、δ_{ki} 和 $\bar{\beta}_{ik}$[13]。

1.2 BRB 推理

BRB 采用 RIMER 方法进行推理,主要包含 3 个步骤:将系统输入值转换为前提属性的个体匹配度,计算每条规则的激活权重以及使用证据推理(Evidential Reasoning, ER)算法合成激活规则。

1.2.1 个体匹配度计算

个体匹配度是指给定的输入值与前提属性的匹配程度。对于式(1)的规则形式,传统的个体匹配度计算可以采用基于效用值的转化方法[14]。对输入的数值 x_i,前提属性 U_i,则 x_i 转化为 U_i 的置信分布形式可以表示如下:

$$S(x_i)=\{(A_{ij},\alpha_{ij});j=1,2,\cdots,J_i\} \tag{2}$$

其中

$$\begin{cases}\alpha_{ij}=\dfrac{A_{i(j+1)}-x_i}{A_{i(j+1)}-A_{ij}}; & A_{ij}\leqslant x_i\leqslant A_{i(j+1)}\\ \alpha_{i(j+1)}=1-\alpha_{ij}\end{cases} \tag{3}$$

$$\alpha_{ik}=0;\quad k=1,2,\cdots,J_i,k\neq j,j+1 \tag{4}$$

其中:J_i 代表第 i 个前提属性的参考等级的个数;A_{ij} 代表第 i 个前提属性的第 j 个参考等级的效用值,且该值会随着参数学习而变化;α_{ij} 代表数值 x_i 转化到前提属性 U_i 第 j 个参考等级上的匹配度。

此时,遍历整个规则库,即可计算得到该输入对每条规则的每个前提属性的个体匹配度 $\alpha^k(x_i,U_i)$,如式(5)所示:

$$\alpha^k(x_i,U_i)=\begin{cases}\alpha_{ij}, & A_i^k=A_{ij}\\ 0, & \text{其他}\end{cases} \tag{5}$$

其中 A_i^k 的定义同式(1)。

1.2.2 激活权重计算

激活权重的计算主要与个体匹配度、规则权重和前提属性权重相关,对于第 k 条规则,假设给定输入 $\boldsymbol{X}=(x_1,x_2,\cdots,x_T)$,则激活权重的计算公式[2]270-271如下所示:

$$w_k(\boldsymbol{X})=\frac{\theta_k*\prod_{i=1}^{T}(\alpha^k(x_i,U_i))^{\bar{\delta}_i}}{\sum_{j=1}^{L}\left[\theta_j*\prod_{i=1}^{T}(\alpha^j(x_i,U_i))^{\bar{\delta}_i}\right]};\quad \bar{\delta}_i=\delta_i/\max_{i=1,2,\cdots,T}\{\delta_i\} \tag{6}$$

其中:θ_k 代表第 k 条规则的规则权重;δ_i 代表第 i 个前提属性的属性权重;$\alpha^k(x_i,U_i)$ 代表输入 x_i 对第 k 条规则前提属性 U_i 的个体匹配度。

1.2.3 ER 算法

在求得每条规则激活权重的基础上,可使用 ER 解析公式[15]合成所有的激活规则。对给定的输入 $\boldsymbol{X}$,结果属性中第 i 个评价等级置信度的合成公式如下:

$$c_{i1}=\prod_{k=1}^{L}\left(w_k(\boldsymbol{X})\beta_{ik}+1-w_k(\boldsymbol{X})\sum_{j=1}^{N}\beta_{jk}\right)-\prod_{k=1}^{L}\left(1-w_k(\boldsymbol{X})\sum_{j=1}^{N}\beta_{jk}\right) \tag{7}$$

$$c_{i2}=\sum_{s=1}^{N}\prod_{k=1}^{L}\left(w_k(\boldsymbol{X})\beta_{sk}+1-w_k(\boldsymbol{X})\sum_{j=1}^{N}\beta_{jk}\right)-(N-1)\prod_{k=1}^{L}\left(1-w_k(\boldsymbol{X})\sum_{j=1}^{N}\beta_{jk}\right)-\prod_{k=1}^{L}(1-w_k(\boldsymbol{X})) \tag{8}$$

$$\beta_i(\boldsymbol{X})=c_{i1}/c_{i2} \tag{9}$$

当 $\sum_{j=1}^{N}\beta_{jk}=1$ 时,式(9)可简化为如下形式:

$$\beta_i(\boldsymbol{X})=\frac{\prod_{k=1}^{L}(w_k(\boldsymbol{X})\beta_{ik}+1-w_k(\boldsymbol{X}))-\prod_{k=1}^{L}(1-w_k(\boldsymbol{X}))}{\sum_{s=1}^{N}\left[\prod_{k=1}^{L}(w_k(\boldsymbol{X})\beta_{sk}+1-w_k(\boldsymbol{X}))-\prod_{k=1}^{L}(1-w_k(\boldsymbol{X}))\right]} \tag{10}$$

合成所有的激活规则后可得到 BRB 系统的分布式置信度:

$$\boldsymbol{f}(\boldsymbol{X})=\{(D_i,\beta_i(\boldsymbol{X})),i=1,2,\cdots,N\} \tag{11}$$

为更直观地表示 BRB 系统的输出,假设结果属性中第 i 个评价等级的效用值为 $\mu(D_i)$,则可推得 BRB 系统数值型输出为:

$$f_U(\boldsymbol{X})=\sum_{i=1}^{N}[\mu(D_i)\beta_i(\boldsymbol{X})] \tag{12}$$

1.3　EBRB

1.3.1　EBRB 的结构表示

在传统 BRB 的基础上,Liu 等[11] 提出了 EBRB。其规则形式表示如下:

$$\begin{aligned} R_k:\text{IF}\ & \{(A_{11}^k,\alpha_{11}^k),(A_{12}^k,\alpha_{12}^k),\cdots,(A_{1J_1}^k,\alpha_{1J_1}^k)\}\ \wedge \\ & \{(A_{21}^k,\alpha_{21}^k),(A_{22}^k,\alpha_{22}^k),\cdots,(A_{2J_2}^k,\alpha_{2J_2}^k)\}\ \wedge\cdots\wedge \\ & \{(A_{T_k1}^k,\alpha_{T_k1}^k),(A_{T_k^2}^k,\alpha_{T_k2}^k),\cdots,(A_{T_kJ_{T_k}}^k,\alpha_{T_kJ_{T_k}}^k)\} \\ \text{THEN}\ & \{(D_1,\bar{\beta}_{1k}),(D_2,\bar{\beta}_{2k}),\cdots,(D_N,\bar{\beta}_{Nk})\}; \\ & \sum_{i=1}^{N}\bar{\beta}_{ik}\leqslant 1 \end{aligned} \tag{13}$$

其中:规则权重为 θ_k;属性权重为 $\delta_{k1},\delta_{k2},\cdots,\delta_{kT_k}$;$k\in\{1,2,\cdots,L\}$。式(13)与式(1)相比,主要区别在于式(13)中每个前提属性的候选值都是置信分布的形式。$A_{ij}^k(i=1,2,\cdots,T_k,j=1,2,\cdots,J_i)$ 代表第 k 条规则的第 i 个前提属性的第 j 个参考等级的效用值,T_k 代表前提属性的个数,J_i 代表第 i 个前提属性的参考等级个数;$\alpha_{ij}^k(i=1,2,\cdots,T_k,j=1,2,\cdots,J_i)$ 代表第 k 条规则的第 i 个前提属性的第 j 个参考等级的置信度;其他变量的定义同式(1)。

1.3.2　EBRB 构建

传统 BRB 构建大多依赖于专家知识,为提高 BRB 系统推理的准确性,同时还需要对 BRB 进行参数训练。然而,参数训练是一个不断迭代的最优化过程,因此大多参数训练方法都比较耗时。为此 Liu 等[11] 在提出 EBRB 的基础上,进一步提出从样本数据中生成 EBRB 的方法。该方法不仅避免对专家知识的依赖,同时还摆脱了耗时的参数训练过程。主要步骤可以归纳如下:

1)从样本数据中选取一部分数据作为生成规则库的数据;

2)确定前提属性和结果属性的评价等级效用值及其个数;

3)将用来生成规则库的输入—输出对转化为与 EBRB 的前提属性和结果属性相同的置信分布形式,如式(13)所示;

4)不断执行步骤 3),直到所有的数据都生成对应的规则;

5)最后,计算每条规则的规则权重。

通过以上 5 步,一个完整的 EBRB 就构建完毕。

1.3.3　EBRB 推理

EBRB 的推理过程是以 1.3.2 节构建的 EBRB 为基础的,其与 BRB 的推理过程类似,都必须经过 3 步,其中激活权重的计算与 ER 合成的方法都与传统的 BRB 一样,即 1.2.2 节和 1.2.3 节所介绍的方法。但由于 EBRB 与 BRB 规则形式上的不同,致使其在个体匹配度的计算上存在差异。以下介绍 EBRB 的个体匹配度的计算方法。具体步骤[11]132 如下:

1)由于 EBRB 的前提属性候选值采用置信分布的形式,如式(13)所示。因此,首先将输入值转化为置信分布形式,具体方法与 1.2.1 节的 BRB 的个体匹配度的计算方法类似。其中主要区别在于此处求得的仅是相对前提属性评价等级的置信分布形式。假设求得的输入 x_i 的分布式类型如下所示:

$$S'(x_i)=\{(A_{ij},\alpha_{ij});j=1,2,\cdots,J_i\} \tag{14}$$

其中:A_{ij} 与式(2)中的 A_{ij} 具有相同的定义,但是由于 EBRB 是不需要进行参数训练的,因此式(14)中的 A_{ij} 在给定后固定不变;α_{ij} 定义同式(2)。

2)当输入值转化为置信分布形式后,Liu 等[11] 首先采用欧几里得距离公式计算两个分布之间的距离,然后再计算个体匹配度。假设输入 x_i 的置信分布形式如式(14)所示,第 i 个前提属性 U_i 的置信分布如下所示:

$$S(U_i)=\{(A_{ij}^k,\alpha_{ij}^k);j=1,2,\cdots,J_i\} \tag{15}$$

其中欧几里得距离计算公式如下:

$$d^k(x_i,U_i)=\sqrt{\sum_{j=1}^{J_i}(\alpha_{ij}-\alpha_{ij}^k)^2} \tag{16}$$

于是,输入 x_i 对前提属性 U_i 的个体匹配度为:

$$\alpha^k(x_i,U_i)=1-d^k(x_i,U_i) \tag{17}$$

1.4　问题提出

Liu 等[11] 提出的 EBRB 的表示、构建与推理方法虽然已被证明具备规则库构建简单、推理效率高的优点,但该方法还存在诸多不完善之处。现实情况中与产生规则库相关的样本数据较多,因而在 EBRB 中会产生大量的规则。此外,当采用 Liu 等[11] 的个体匹配度计算方法及确定激活规则的策略时,EBRB 中的大部分规则会被激活。然而在激活规则中大多数规则对推理结果往往是不起作用的,甚至会降低推理准确性。尤其当样本数据中存在噪声数据时,对 EBRB 推理准确性的影响较为明显。换言之,Liu 等[11] 方法中存在不合理的规则激活机制,导致推理结果不够准确。针对上述问题,本文对 Liu 等[11] 方法进行改进,以获得一种更加合理、有效的规则库构建及推理方法。

2　数据驱动的 BRB 构建方法

针对目前 EBRB 构建方法未能使规则库系统具有理想的推理准确性,本文在 EBRB 构建方法的基础上,提出一种数据驱动的 BRB 构建方法。以下将具体介绍该方法。

2.1　生成规则库的前件和后件

规则库的前件是由输入数据直接作为前提属性的参考值。假设一个函数 $z=g(x,y)$ 及一个输入数据 $(x,y,z)=(1,2,3)$,则由该数据生成的规则前件为 IF $x=1\wedge y=2$。

对于规则库的后件,采用文献[14]的基于效用值的转化方法获取。假设输入的数值 x_i,结果属性 C_i,则 x_i 转化为 C_i 的置信分布形式可以表示如下:

$$S(x_i)=\{(D_{ij},\alpha_{ij});j=1,2,\cdots,N\} \tag{18}$$

其中

$$\alpha_{ij}=\frac{D_{i(j+1)}-x_i}{D_{i(j+1)}-D_{ij}};\quad \alpha_{i(j+1)}=1-\alpha_{ij},D_{ij}\leqslant x_i\leqslant D_{i(j+1)} \tag{19}$$

$$\alpha_{ik}=0;\quad k=1,2,\cdots,N,k\neq j,j+1 \tag{20}$$

其中:N 代表结果属性的评价等级的个数,D_{ij} 代表第 i 个结果属性的第 j 个评价等级的效用值,α_{ij} 代表数值 x_i 转化到结果属性 C_i 第 j 个评价等级上的置信度。

2.2　规则权重计算

由于样本数据中可能存有噪声数据,为减小噪声数据所产生的错误推理,因此需要使用规则权重区分不同规则间的重要程度。针对该情况,本文在 Liu 等[11] 方法的基础上给出一种适合 BRB 规则形式的规则权重计算方法。

2.2.1　相似性度量

规则相似性包含前件相似性(Similarity measures of Rule

Antecedent, SRA)和后件相似性(Similarity measures of Rule Consequent, SRC)[16]261。首先,在计算 SRA 和 SRC 之前,先讨论相似性度量的方法。

假设有两个数值 x 和 y,它们之间的距离可通过以下公式来计算:

$$d(x,y) = |x-y| \tag{21}$$

为保证下文的相似性大小在0到1之间,则必须对距离进行归一化,如下所示:

$$d(x,y) = d(x,y)/m \tag{22}$$

其中正常数 m 的大小根据具体应用而定。于是,最终的距离计算公式可定义如下:

$$d(x,y) = \begin{cases} |x-y|/m, & |x-y| \leqslant m \\ 1, & \text{其他} \end{cases} \tag{23}$$

那么,可得相似性的计算公式:

$$Sim(x,y) = 1 - d(x,y) \tag{24}$$

2.2.2 前件和后件相似性

假设存在两条规则(其中结果属性未转化为分布的原始数值形式):

$$R_i: \text{IF } U_1 = A_1^i \wedge U_2 = A_2^i \wedge \cdots \wedge U_T = A_T^i, \text{THEN } V = B_i \tag{25}$$

$$R_k: \text{IF } U_1 = A_1^k \wedge U_2 = A_2^k \wedge \cdots \wedge U_T = A_T^k, \text{THEN } V = B_k \tag{26}$$

其中:T 代表规则前提属性的个数;A_i^j 代表前提属性的候选值(就本文构建的规则库而言,A_i^j 是构成第 j 条规则的数据中对应前提属性 U_i 的数值);B_i 代表第 i 条规则结果属性未转化为置信分布的候选值,即构成第 i 条规则的数据对应结果属性 V 的数值。

那么,上述两条规则的前件相似性可以定义如下:

$$SRA(i,k) = \sum_{j=1}^{T} [\delta_j * Sim(A_j^i, A_j^k)] \tag{27}$$

后件相似性定义如下:

$$SRC(i,k) = Sim(B_i, B_k) \tag{28}$$

其中:式(27)的 $\delta_j (j = 1,2,\cdots,T)$ 代表第 j 个前提属性的属性权重;$Sim(A_j^i, A_j^k)$ 代表第 i 条规则和第 k 条规则的第 j 个前提属性之间的相似性,由式(24)计算得到。而式(28)的 $Sim(B_i, B_k)$ 代表第 i 条规则和第 k 条规则的结果属性的相似性,同理由式(24)计算得到(这里假设规则只有一个结果属性)。

2.2.3 规则的一致度

利用以上计算得到的前件和后件相似性,可计算两两规则之间的一致度。具体计算公式[17]如下:

$$Cons(R_i, R_k) = \exp\{-(SRA(i,k)/SRC(i,k) - 1.0)^2/(1/SRA(i,k))^2\} \tag{29}$$

其中 $Cons(R_i, R_k) \in [0,1]$。

根据两两规则的一致度,计算两两规则的不一致度:

$$Uncons(R_i, R_k) = 1 - Cons(R_i, R_k) \tag{30}$$

第 i 条规则对于整个规则库的不一致度如下:

$$Incons(i) = \sum_{k=1, k\neq i}^{L} Uncons(R_i, R_k) \tag{31}$$

即可计算规则库的整体不一致度[16]262,如下所示:

$$\xi_{\text{Incons}} = \sum_{i=1}^{L} Incons(i) \tag{32}$$

通过以上对每条规则的不一致度和规则库的整体不一致度的计算,可得规则权重的计算公式,如下所示:

$$\theta_k = 1 - \lambda * Incons(i)/\xi_{\text{Incons}} \tag{33}$$

其中常数 λ 的大小根据具体的应用而定。

至此,一个完整的规则库构建完毕。通过以上介绍可以发现,规则数 L 为生成规则的数据个数。

3 引入 80/20 法则的 BRB 推理方法

3.1 与传统 BRB 的区别

虽然本文由样本数据生成的规则在结构上与传统由专家给定的规则都满足式(1)的形式,但本文构建规则的前提属性的候选值与传统的 BRB 仍然是有所区别的。下面举例说明。

对传统的由专家构建的 BRB 必须事先给定前提属性的参考等级。比如输油管道检漏实验中的输入和输出的流量差(Flow Differential, FD)属性须事先设定 8 个参考等级[3]106,语义值为:

$$\{\text{NL}, \text{NM}, \text{NS}, \text{NVS}, \text{Z}, \text{PS}, \text{PM}, \text{PL}\} \tag{34}$$

对应的量化值为:

$$\{-10, -5, -3, -1, 0, 1, 2, 3\} \tag{35}$$

根据专家知识构建规则的前提属性 FD 的候选值就局限在这 8 个参考等级上。即便通过候选值训练之后,新得到的置信规则库的候选值在数值上会发生变化,但是对整体规则库来讲,所有的前提属性的不同候选值的个数依然不变。然而,本文提到的由样本数据构建的置信规则库的规则 FD 前提属性的候选值并不是选自这 8 个参考等级,而是来源于样本数据的 FD 属性值。因此,当样本数据个数较多时,通过本文第 2 章介绍的规则库构建方法构建出来的规则条数也比较多,那么规则的前提属性也将存在很多不同数值的候选值,并且毫无规律。于是,本文在传统 BRB 的推理基础上,提出了引入 80/20 法则的规则推理方法。

3.2 个体匹配度计算

由于本文构建的规则库的特殊性,以下给出一种针对该类规则库的个体匹配度计算方法。假设输入 x_i,前提属性 U_i,则 x_i 对 U_i 的个体匹配度计算公式如下:

$$\alpha^k(x_i, U_i) = \begin{cases} 1 - |x_i - A_i^k|/m, & |x_i - A_i^k| \leqslant m \\ 0, & \text{其他} \end{cases} \tag{36}$$

其中:A_i^k 代表第 k 条规则的第 i 个前提属性的候选值;m 是一个常数,取值同式(23)。

3.3 置信规则库推理过程

在由本文第 2 章构建出来的置信规则库基础上,以下介绍该种类型的置信规则库的推理方法。具体步骤如下:

1)对给定一个输入 $\boldsymbol{X} = (x_1, x_2, \cdots, x_T)$,首先根据式(36)可以计算得到第 k 条规则中该输入对每个前提属性 U_i 的个体匹配度为:

$$\boldsymbol{\rho}^k = \{\alpha^k(x_1, U_1), \alpha^k(x_2, U_2), \cdots, \alpha^k(x_T, U_T)\} \tag{37}$$

2)不断执行步骤1),直到输入 $\boldsymbol{X}$ 对所有规则的每个前提属性的个体匹配度都求完为止。

3)根据步骤1)得到的如式(37)所示的个体匹配度,并结合通过 2.2 节的方法计算得到的规则权重,再利用式(6)计算输入 $\boldsymbol{X}$ 对所有规则的激活权重,如下所示:

$$\boldsymbol{W}(\boldsymbol{X}) = \{w_1(\boldsymbol{X}), w_2(\boldsymbol{X}), \cdots, w_L(\boldsymbol{X})\} \tag{38}$$

4) 根据 80/20 法则,选择激活前 t 条规则(具体激活方法见 3.4 节),如下所示:

$$\boldsymbol{W}'(\boldsymbol{X}) = \{w'_1(\boldsymbol{X}), w'_2(\boldsymbol{X}), \cdots, w'_t(\boldsymbol{X})\} \tag{39}$$

5)利用 ER 算法将通过步骤 4)被激活的 t 条规则进行合成,可得到 BRB 系统输出。该输出可分为式(11)或(12)两种形式。

3.4　引入 80/20 法则的规则激活机制

通常情况下,本文构建的规则库会存在大量的规则。而当规则条数较多时,通过 3.3 节的步骤 3)将得到过多的激活规则。然而,对于大部分系统而言,一个给定输入所对应的有用的规则通常只占少部分。同时,由于样本数据中难免存有噪声数据,所以从样本数据中生成的规则库也必然存有噪声规则。虽然本文在第 2 章通过引入一致性测度的方法计算每条规则的规则权重,降低了噪声规则的规则权重,但是若采用 Liu 等[11]的规则激活方法,噪声规则被激活的概率依然较大,进而影响到推理结果的准确性。

鉴于以上原因,本文提出了一种引入 80/20 法则的规则激活机制。即对给定一个输入,只激活对输入最重要的那些规则,可以是 20%,也可以少于 20%,或者多于 20%(以下都用 $r\%$ 代替),这些都根据具体的应用而定。其中,"对输入最重要的那些规则"是指式(38)中的所有规则激活权重(大于 0)最大的前 $r\%$ 条规则。具体的步骤如下:

1)对式(38)中的所有规则的激活权重进行从大到小排序。表示如下:

$$\boldsymbol{W}^s(\boldsymbol{X}) = \{w_1^s(\boldsymbol{X}), w_2^s(\boldsymbol{X}), \cdots, w_L^s(\boldsymbol{X})\} \tag{40}$$

其中 $w_1^s(\boldsymbol{X}) \geqslant w_2^s(\boldsymbol{X}) \geqslant \cdots \geqslant w_L^s(\boldsymbol{X})$。

2)取出式(40)中激活权重大于 0 的规则,假设有 n 条,如下所示:

$$\boldsymbol{W}^n(\boldsymbol{X}) = \{w_1^s(\boldsymbol{X}), w_2^s(\boldsymbol{X}), \cdots, w_n^s(\boldsymbol{X})\} \tag{41}$$

3)取出式(41)的 n 条规则中的前 $r\%$ 条规则,即有 $n * r\% = t$。然后再对 t 条规则的激活权重重新进行归一化处理,最终得到式(39)所示的 t 条规则的激活权重。同时,将 n 条规则中不属于 $r\%$ 之内的规则的激活权重置为 0。

通过以上 3 步,可得到一个新的每条规则的激活权重,接着,即可通过证据推理方法将被激活的规则进行合成。本文第 4 章的实验可以验证引入 80/20 法则的激活机制可以提高 BRB 系统的推理精度。

4　实验结果与分析

为验证本文方法的有效性,以安装在英国的一条 100 多公里长的输油管道作为研究对象,并且采用该输油管道的真实泄露数据作为测试数据。

当输油管道系统正常工作时,如果油液的输入量大于输出量,那么管道中的油液总量变大,这样就会导致油液对管道产生的压力也会变大。但是,如果油液的输入量大于输出量,而油液对管道产生的压力却减小的话,那么可以说明该管道很可能发生了泄露。于是,采用输入与输出的流量差(FD)、油液对管道产生的平均压力差(Pressure Differential, PD)以及泄露大小(Leak Size, LS)作为泄露数据来实现对管道泄露的检测和泄露大小的估计。

4.1　实验数据和环境

实验数据来自于文献[3]中用到的 2 008 组管道泄露数据。该 2008 组数据是以每 10 s 为采样周期收集到的从正常到发生 25% 泄露时的数据。本实验的实验环境为:AMD Athlon II X4 645 Processor 3.10 GHz;4 GB 内存;Windows 7 操作系统;算法由 Visual C ++ 6.0 和 Matlab R2013a 编写。

4.2　规则库构建

本文给出的规则库构建方法无需事先给定前提属性的参考等级。但是,需要预先给定结果属性 LS 的评价等级,具体定义如下:

$$\boldsymbol{D} = \{d_1, d_2, d_3, d_4, d_5\} = \{\mathrm{Z}, \mathrm{VS}, \mathrm{M}, \mathrm{H}, \mathrm{VH}\} \tag{42}$$

其中,这些语义值对应的量化值如下:

$$\boldsymbol{D} = \{0, 2, 4, 6, 8\} \tag{43}$$

为构建规则库,本文采用从 2008 组样本数据中间隔选取数据的方式。这里间隔选取是指从该 2008 组样本数据中每隔几个点选取一个数据(即以一定步长选取数据)作为生成规则库的数据。这里选择步长为 4,则生成 502 组数据。然后任意删除两个点,得到 500 组数据。最后根据本文第 2 章介绍的 BRB 构建方法,将得到由 500 条规则构成的规则库。

4.3　有效性验证

本文的主要工作在于提出了一种基于 BRB 结构的规则库构建方法以及引入 80/20 法则的规则激活机制。接下来将从实验验证所提方法的优越性。

本文中存在两个参数:1)2.2.1 节的距离度量的常数 m;2)3.4 节推理所取规则条数的百分比 $r\%$。其中由于本实验中存在 3 个属性 FD、PD、LS,所以可用一个三维向量 $\boldsymbol{M}$ 来表示不同属性下 m 的不同取值。根据经验,$\boldsymbol{M}$ 可取(1, 0.01, 2)。至于第 2 个参数,以下给出该参数变化下的实验结果。为说明实验结果的优劣,这里引入平均绝对误差(Mean Absolute Error, MAE)作为误差衡量标准。如表 1,给出了 $r\%$ 从 10% 到 100% 变化下的 MAE,其中:Liu_EBRB 代表 Liu 等[11]的 EBRB 推理方法,DD_BRB 代表本文的推理方法,并采用2008 组样本数据作为测试数据。

表 1　两种方法在不同 r 变化下的 MAE 值

r	MAE		r	MAE	
	DD_BRB	Liu_EBRB		DD_BRB	Liu_EBRB
10	0.179 82	0.211 92	60	0.183 80	0.238 13
20	0.173 42	0.219 61	70	0.186 37	0.238 63
30	0.174 15	0.222 56	80	0.188 21	0.239 23
40	0.176 60	0.233 41	90	0.190 26	0.240 25
50	0.180 27	0.239 70	100	0.191 85	0.240 26

由于 Liu_EBRB 方法本身未引入 80/20 法则,为了便于更有效地进行实验对比,故对 Liu_EBRB 方法也引入 80/20 法则。

为了更直观地说明引入 80/20 法则的规则激活方法的有效性,以下绘制 r 变化下的 MAE 曲线,如图 1 所示。

从表 1 和图 1 可以得出两个结论:

1)在未引入 80/20 法则的前提下,本文构建出来的基于 BRB 结构的规则库在推理精度上优于 Liu 等[11]的基于 EBRB 结构的规则库。对表 1 和图 1 而言,当 $r = 100$ 时,意味着此时的规则激活方法未引入 80/20 法则。

2)不论是 Liu 等[11]的 EBRB 还是本文构建出来的规则

库,当引入 80/20 法则后,规则库的推理精度都将提高。

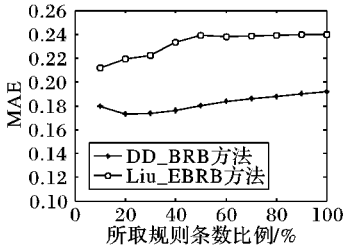

图 1 两种方法在不同 r 变化下的 MAE 曲线

由此可见,引入 80/20 法则后的本文构建出来的 BRB 系统比 Liu 等[11]的 EBRB 系统有更高的推理精度。

从表 1 可以看出,当 $r = 20$ 时,DD_BRB 的 MAE 值最小,可达到 0.173 42。如图 2,给出了当 $r = 20$ 时,2 008 组测试数据下的泄露大小的真实值和估计值。从图 2 可发现,除个别点之外,大部分点都有较好的拟合效果。

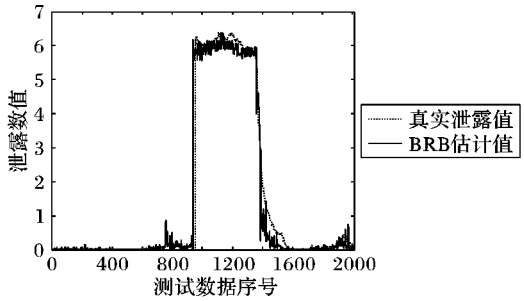

图 2 2 008 组测试数据下的 BRB 系统的估计输出和真实输出

4.4 效率及误差对比

为了更进一步验证本文提出的方法的性能,以下将本文方法与 Liu 等[11]的方法及 Chen 等[13]的方法进行效率和误差对比。效率上,分别从生成规则库的时间和测试时间两部分进行讨论,当二者之和越小,则效率越高。误差上,讨论 MAE 和均方根误差(Root Mean Square Error, RMSE)两种误差衡量标准,当误差越小,则系统的精度越高。其中:效率和误差对比如表 2 所示。这里,Chen 等[13]的方法为 Chen_BRB,其中 Chen 等[13]的方法用到 Matlab 的 FMINCON 函数进行训练,故而 Chen_BRB 方法的耗时包括训练时间和测试时间。

表 2 3 种方法的耗时和误差对比

方法	耗时/s	MAE	RMSE
DD_BRB	0.83	0.173 42	0.555 56
Liu_EBRB	0.91	0.240 26	0.731 71
Chen_BRB	2 838.62	0.211 99	0.585 09

其中,对 Chen_BRB 方法而言,500 组间隔数据是作为规则库的训练数据。而对 DD_BRB 方法以及 Liu_EBRB 方法均是将 500 组数据作为生成规则库的数据。然后用上文提到的 2 008 组样本数据作为测试数据对这 3 种方法的性能进行测试。同时,对 DD_BRB 方法进行验证时,取 3.4 节中推理所取规则条数的百分比 $r\%$ 为 20%。

从表 2 可看出,本文提出的方法虽然在效率上与 Liu 等[11]的方法相差无几,但是在精度上明显优于 Liu 等[11]的方法。同样可以发现,本文方法比 Chen 等[13]的方法不仅效率高很多,而且在精度上也优于 Chen 等[13]的方法。同时,通过这个实验可以发现规则库参数训练过程比较耗时,而本文提出的方法与 Liu 等[11]的方法都无需这个过程,故而在效率上有明显的优势。总之,通过比较,可以验证利用本文方法得到的置信规则库系统具有良好的性能。

5 结语

Liu 等[11]提出的 EBRB 的表示、构建和推理方法在效率上较传统的方法有较大的提升。然而,该 EBRB 系统推理精度不高,且易受噪声数据干扰。因此,本文首先将 Liu 等[11]的 EBRB 构建方法应用在 BRB 的构建上,并且针对 BRB 的规则形式作了适当的改进;同时提出了引入 80/20 法则的 BRB 推理方法;最后,通过输油管道检漏实例验证了本文提出的方法不仅是在效率上,还是在精度上都是有效可行的。然而,本文的研究还处于初步阶段,对于本文提出的数据驱动的规则库构建方法还有许多亟待研究的问题,如:当存在大规模的样本数据集时,如何从中选取用来生成规则库的数据将是下一步研究工作的重点。

参考文献:

[1] PEARL J. Probabilistic reasoning in intelligent systems [M]. Waltham: Morgan Kaufmann Publishers, 1988.

[2] YANG J-B, LIU J, WANG J, *et al.* Belief rule-base inference methodology using the evidential reasoning approach-RIMER [J]. IEEE Transactions on System, Man and Cybernetics, Part A: Systems and Humans, 2006, 36(2): 266 – 285.

[3] XU D-L, LIU J, YANG J-B, *et al.* Inference and learning methodology of belief-rule-based expert system for pipeline leak detection [J]. Expert Systems with Applications, 2007, 32(1): 103 – 113.

[4] WANG Y-M, YANG J-B, XU D-L, *et al.* Consumer preference prediction by using a hybrid evidential reasoning and belief rule-based methodology [J]. Expert Systems with Applications, 2009, 36(4): 8421 – 8430.

[5] LI B, WANG H-W, YANG J-B, *et al.* A belief-rule-based inventory control method under nonstationary and uncertain demand [J]. Expert Systems with Applications, 2011, 38(12): 14997 – 15008.

[6] KONG G-L, XU D-L, BODY R, *et al.* A belief rule-based decision support system for clinical risk assessment of cardiac chest pain [J]. European Journal of Operational Research, 2012, 219(3): 564 – 573.

[7] CHEN Y-W, POON S-H, YANG J-B, *et al.* Belief rule-based system for portfolio optimisation with nonlinear cash-flows and constraints [J]. European Journal of Operational Research, 2012, 223(3): 775 – 784.

[8] YANG J-B, LIU J, XU D-L, *et al.* Optimization models for training belief-rule-based systems [J]. IEEE Transactions on System, Man and Cybernetics, Part A: Systems and Humans, 2007, 37(4): 569 – 585.

[9] CHANG R, WANG H, YANG J. An algorithm for training parameters in belief rule-bases based on the gradient and dichotomy methods [J]. Systems Engineering, 2007, 25(S): 287 – 291. (常瑞, 王红卫, 杨剑波. 基于梯度法和二分法的置信规则库参数训练方法 [J]. 系统工程, 2007, 25(增刊): 287 – 291.)

[10] ZHOU Z-G, LIU F, JIAO L-C, *et al.* A bi-level belief rule based decision support system for diagnosis of lymph node metastasis in gastric cancer [J]. Knowledge-Based Systems, 2013, 54: 128 – 136.

(下转第 2169 页)

利用所提出的指标,可以自动获得最佳聚类数,实现聚类的无监督学习过程。所提出的指标的优点是:1)克服了 FCM 算法中需要预先指定类数的缺点;2)克服了在使用距离度量的情况下,由于数据分布的不均匀性导致计算模糊程度不准确的缺点。实验结果表明了所提出指标的有效性,且所提出的指标对模糊子具有较好的鲁棒性。由于数据分布特征的多样性,如何定义一个更好的有效性指标,快速地发现最符合数据自然分布的聚类结果,仍是一个值得深入研究的问题。

参考文献:

[1] REZAEE B. A cluster validity index for fuzzy clustering [J]. Fuzzy Sets and Systems, 2010, 161(23): 3014 – 3025.

[2] ZHANG Y J, WANG W N, ZHANG X, *et al.* A cluster validity index for fuzzy clustering [J]. Information Sciences, 2008, 178(4): 1205 – 1218.

[3] ZALIK K R. Cluster validity index for estimation of fuzzy clusters of different sizes and densities [J]. Pattern Recognition, 2010, 43(10): 3374 – 3390.

[4] KIMA D W, LEE K H, LEE D. On cluster validity index for estimation of the optimal number of fuzzy clusters [J]. Pattern Recognition, 2004, 37(10): 2009 – 2025.

[5] WANG W, ZHANG Y. On fuzzy cluster validity indices [J]. Fuzzy Sets and Systems, 2007, 158(19): 2095 – 2117.

[6] YUE S, WANG J, WU T, *et al.* A new separation measure for improving the effectiveness of validity indices [J]. Information Sciences, 2010, 180(5): 748 – 764.

[7] HUANG K. Applications of an enhanced cluster validity index method based on the fuzzy C-means and rough set theories to partition and classification [J]. Expert Systems with Applications, 2010, 37(12): 8757 – 8769.

[8] MASSON M, WU T. ECM: an evidential version of the fuzzy C-means algorithm [J]. Pattern Recognition, 2008, 41(4): 1384 – 1397.

[9] WU K, YANG M. Robust cluster validity indexes [J]. Pattern Recognition, 2009, 42(11): 2541 – 2550.

[10] SUN X, ZHAO Y, WANG H-L, *et al.* Sensitivity of digital soil maps based on FCM to the fuzzy exponent and the number of clusters [J]. Geoderma, 2012, 171/172: 24 – 34.

[11] MITRA S, PEDRYCZ W, BARMAN B. Shadowed C-means: integrating fuzzy and rough clustering [J]. Pattern Recognition, 2010, 43(4): 1282 – 1291.

[12] DUNN J C. Well-separated clusters and the optimal fuzzy partitions [J]. Journal of Cybernetics, 1974, 4(1): 95 – 104.

[13] BEZDEK J C. Pattern recognition with fuzzy objective function algorithms [M]. Norwell: Kluwer Academic Publishers, 1981.

[14] BEZDEK J C. Cluster validity with fuzzy sets [J]. Journal of Cybernetics, 1974, 3(3): 58 – 73.

[15] BEZDEK J C. Numerical taxonomy with fuzzy sets [J]. Journal of Mathematical Biology, 1974, 1(1): 57 – 71.

[16] DAVE R N. Validating fuzzy partitions obtained through C-shells clustering [J]. Pattern Recognition Letters, 1996, 17(6): 613 – 623.

[17] WINDHAM M P. Cluster validity for the fuzzy C-means clustering algorithm [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1982, 4(4): 357 – 363.

[18] FUKUYAMA Y, SUGENO M. A new method of choosing the number of clusters for the fuzzy C-means method [C]// Proceedings of the 5th Fuzzy Systems Symposium. Kobe: [s. n.], 1989: 247 – 250.

[19] GUNDERSON R. Applications of fuzzy ISODATA algorithms to star-tracker printing systems [C]// Proceedings of the 7th Triennial World IFAC Congress. Helsinki: [s. n.], 1978: 1319 – 1323.

[20] DAVIES D L, BOULDIN D W. A cluster separation measure [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1979, 1(2): 224 – 227.

[21] GATH I, GEVA A B. Fuzzy clustering for the estimation of the parameters of the components of mixtures of normal distributions [J]. Pattern Recognition Letters, 1989, 9(2): 77 – 86.

[22] XIE X, BENI G. A validity measure for fuzzy clustering [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1991, 13(4): 841 – 847.

[23] WINKLER R, KLAWONN F, KRUSE R. Fuzzy C-means in high dimensional spaces [J]. International Journal of Fuzzy System Applications, 2011, 1(1): 1 – 16.

[24] ZAHID N, LIMOURI M, ESSAID A. A new cluster validity for fuzzy clustering [J]. Pattern Recognition, 1999, 32(7): 1089 – 1097.

[25] WU K, YANG M. A cluster validity index for fuzzy clustering [J]. Pattern Recognition Letters, 2005, 26(9): 1275 – 1291.

[26] PAKHIRA M K, BANDYOPADHYAY S, MAULIK U. Validity index for crisp and fuzzy clusters [J]. Pattern Recognition, 2004, 37(3): 487 – 501.

(上接第 2160 页)

[11] LIU J, MARTINEZ L, CALZADA A, *et al.* A novel belief rule base representation, generation and its inference methodology [J]. Knowledge-Based Systems, 2013, 53: 129 – 141.

[12] HU Y. The application of the 80-20 rule in the evaluation of key performance indicators [J]. Construction Machinery and Maintenance, 2009(5): 116 – 117. (胡玉美. 二八法则在关键绩效指标考评中的应用[J]. 工程机械与维修, 2009(5): 116 – 117.)

[13] CHEN Y-W, YANG J-B, XU D-L, *et al.* Inference analysis and adaptive training for belief rule based systems [J]. Expert Systems with Applications, 2011, 38(10): 12845 – 12860.

[14] YANG J-B. Rule and utility based evidential reasoning approach for multiattribute decision analysis under uncertainties [J]. European Journal of Operational Research, 2001, 131(1): 31 – 61.

[15] CHEN Y-W, YANG J-B, XU D-L, *et al.* On the inference and approximation properties of belief rule based systems [J]. Information Sciences, 2013, 234: 121 – 135.

[16] LIU J, MARTINEZ L, RUAN D, *et al.* Optimization algorithm for learning consistent belief rule-base from examples [J]. Journal of Global Optimization, 2011, 51(2): 255 – 270.

[17] JIN Y, von SEELEN W, SENDHOFF B. On generating FC^3 fuzzy rule systems from data using evolution strategies [J]. IEEE Transactions on System, Man and Cybernetics, Part B: Cybernetics, 1999, 29(6): 829 – 845.