DOI:10.7544/issn1000-1239.20200702
59(3):661-673,2022

数据缺失的扩展置信规则库推理方法

刘永裕[1] 巩晓婷[2] 方炜杰[1] 傅仰耿[1]
[1](福州大学数学与计算机科学学院 福州 350116)
[2](福州大学决策科学研究所 福州 350116)
(liuyoyu@outlook.com)

Extended Belief Rule Base Reasoning Approach with Missing Data

Liu Yongyu[1], Gong Xiaoting[2], Fang Weijie[1], and Fu Yanggeng[1]
[1](*College of Mathematics and Computer Science, Fuzhou University, Fuzhou* 350116)
[2](*Institute of Decision Sciences, Fuzhou University, Fuzhou* 350116)

Abstract The data-driven constructed extended belief rule-based system can deal with uncertainty problems with both quantitative data and qualitative knowledge. It has been widely researched and applied in recent years, but infrequently been involved in the field of incomplete data. This study conducts research focusing on the performance of the extended belief rule-based system applied to incomplete datasets and proposes a novel reasoning approach for the case of data missing. First, a disjunctive extended rule base is constructed and the optimal number of antecedent attribute referential values is discussed through validation experiments. Then a method for generating a disjunctive belief rule base from incomplete data and consisting of disjunctive belief rule base is proposed, and an attenuation factor is introduced to modify the weight of incomplete rules to make the aggregation of information more reasonable. Finally, this paper conducts experiments on several commonly used datasets selected from UCI to validate the improvement of the proposed method. The experiments are designed with various degrees and patterns of data missing, and the performance of the improved system is analyzed and compared with some conventional mechanisms. Experimental comparison with other methods shows that while the new method performs well on complete datasets, it also shows better and more stable inference effects on datasets with different degrees of missing and patterns.

Key words belief rule base; evidence reasoning; data driven; data missing; incomplete dataset

摘 要 数据驱动的扩展置信规则库专家系统能够处理含有定量数据或定性知识的不确定性问题.该方法已被广泛地研究和应用,但仍缺乏在不完整数据问题上的研究.鉴于此,针对不完整数据集上的问题,提出一种新的扩展置信规则库专家系统推理方法.首先提出基于析取范式的扩展规则结构,并通过实验讨论了在新的规则结构下,置信规则前提属性参考值个数对推理方法的影响;然后提出通过不完整数据生成具有不完整置信规则,并构成析取范式置信规则库的方法,同时引入衰减因子修正不完整规则权重,使不完整规则可以更合理地参与信息融合过程;最后,选取若干个公共数据集对所提方法进行验证.

收稿日期:2020-09-08;**修回日期**:2021-01-04
基金项目:国家自然科学基金项目(61773123);福建省自然科学基金项目(2019J01647);工信部工业互联网创新发展工程项目(TC19083WB)
This work was supported by the National Natural Science Foundation of China (61773123), the Natural Science Foundation of Fujian Province of China (2019J01647), and the Industrial Internet Innovation and Development Project of the Ministry of Industry and Information Technology, China (TC19083WB).
通信作者:傅仰耿(ygfu@qq.com)

与其他方法的实验对比显示,新方法在完整数据集上有良好表现的同时,对具有不同缺失程度和缺失模式的数据集表现出更好更稳定的推理效果.

关键词 置信规则库;证据推理;数据驱动;数据缺失;不完整数据集

中图法分类号 TP18

现今,数据收集渠道呈现多样化,人们可以通过在线调查、传感器和测量等各种方式收集数据,然而在大部分的工业和商业领域的数据采集过程中,不可避免地会由各种原因造成数据缺失,例如测量不正确、数据采集问题、传感器故障、数据集中遗漏录入等情况.数据的缺失在机器学习、数据挖掘和信息系统中容易造成低效的数据分析和不准确的决策结果,因此数据缺失问题是数据驱动模型需要解决关键问题.

目前对缺失数据的处理方式主要分 3 种情况:1)删除数据集中含有缺失值的数据,但该方法不适用于缺失率高的数据,同时删除数据会丢失该条数据中其他有用的信息.2)采用统计学或机器学习方法利用已有信息对缺失项进行填补,例如手动填补、用均值或众数填补、基于 K 近邻的填补[1-3]、基于期望最大化的填补[4-5]等.K 近邻填补算法能够很好地对数据的缺失项进行填补,但它需要搜寻整个数据集来寻找合适的 K 个近邻,当数据量过大时,这个过程会非常耗时;期望最大化填补只考虑了定量类型的数据,没办法对定性数据进行填补.3)对属性值缺失的数据直接使用的方法,如基于模糊 C 均值的方法[6]、基于 SVM(support vector machine)的方法[7]、基于贝叶斯的方法[8-9]等;但这些方法也或多或少存在局限性,例如基于模糊 C 均值和基于 SVM 方法需要假设缺失数据服从某种分布,但在现实中该数据的分布往往是不可知的.

Yang 等人[10]在 2006 年提出基于证据理论的置信规则库推理方法(belief rule-base inference methodology using the evidential reasoning approach, RIMER),该方法将 D-S 证据理论、贝叶斯概率理论、模糊集理论和传统的 IF-THEN 规则库系统相结合,使其具有能够处理信息不确定性的情况.目前该系统已经广泛运用于消费者偏好预测[11-13]、风险评估[14-16]和医疗诊断[17-19]等领域.

Liu 等人[20]在置信规则库(belief rule base, BRB)系统的基础上提出了扩展置信规则库(extended belief rule base, EBRB)系统,EBRB 能够通过历史数据直接生成置信规则,避免了繁复的 BRB 系统参数优化过程,且解决了规则数"组合爆炸"的问题.该系统目前已经成功应用于健康评估[21]、环境治理成本预测[22]等领域[23-24].

在上述研究的基础上,针对 BRB 系统的数据缺失问题,鱼蒙等人[25]采用抽样思想,通过历史数据获取样本分布并对样本分布分层抽样填补缺失项,再使用 2 次证据推理(evidential reasoning, ER)算法生成最终结果,但该方法不适用于含有噪声或离群点的数据集;王小燕等人[26]通过建立多个不完备子 BRB 系统并合成完备的 DBRB 系统,该方法虽然降低系统的复杂性,但最后推理精度有所下降;Chang 等人[27]借鉴自组织映射思想,用不完整数据训练多个子置信规则库并合成最终析取范式置信规则库,但该方法对数据缺失模式要求比较高,对于缺失率较高的数据,系统复杂度增大;Li 等人[28]使用 BRB 通过历史数据合成缺失项数据对数据集进行补全,并应用在医疗评估中,但该方法面对多属性缺失或缺失率较高的数据集时,仍无法保证最后的推理准确率;Liang 等人[29]用历史数据建立模糊关联规则来解决数据不完备的问题,为解决输入数据缺失提供了思路,但该方法只能处理输入数据不完整的情况,无法处理含有缺失值的数据集.

目前对数据缺失或不完备问题已有一些解决方法,但方法都有自身不足的地方,所以为了解决该问题,本文提出能够处理数据缺失的扩展置信规则库方法,主要贡献有 2 个方面:

1) 提出基于析取范式的扩展置信规则形式,并说明了在该形式下 EBRB 系统的规则前提属性参考值个数的选择会对系统的推理性能有明显影响.

2) 利用不完整数据生成不完整规则,同时引入衰减因子调节不完整规则的权重,在不提高系统复杂性的前提下,解决 EBRB 系统中的数据缺失问题.

1 EBRB 专家系统

1.1 规则表示

$$R_k:\text{IF } U_1 \text{ is } \{(A_{11},\alpha_{11}^k),(A_{12},\alpha_{12}^k),\cdots,(A_{1J_1},\alpha_{1J_1}^k)\} \wedge U_2 \text{ is } \{(A_{21},\alpha_{21}^k),$$

$$(A_{22},\alpha_{22}^k),\cdots,(A_{2J_2},\alpha_{2J_2}^k)\}\wedge\cdots\wedge$$
$$U_T \text{ is } \{(A_{T1},\alpha_{T1}^k),(A_{T2},\alpha_{T2}^k),\cdots,$$
$$(A_{TJ_T},\alpha_{TJ_T}^k)\}$$
$$\text{THEN } \{(D_1,\beta_1^k),(D_2,\beta_2^k),\cdots,(D_N,\beta_N^k)\}$$

其中,θ_k 表示规则权重,$\delta_1,\delta_2,\cdots,\delta_T$ 表示属性权重;

$$\beta\geqslant 0,\ \alpha_{ij}^k\geqslant 0,\ \sum_{j=1}^{N}\beta_j^k\leqslant 1,\ \sum_{j=1}^{J_i}\alpha_{ij}^k\leqslant 1. \quad (1)$$

规则 R_k 的符号说明[20]:k 表示第 k 条规则,U_i 表示第 i 个前提属性;A_{ij} 表示第 i 个属性的第 j 个前提属性参考值,α_{ij} 表示第 i 个属性第 j 个属性参考值的置信度,(A_{ij},α_{ij}) 表示属性 U_i 的置信分布形式;T 表示前提属性个数;J_i 表示第 i 个前提属性的参考值个数;D_i 表示第 i 个评价等级,β_i^k 表示第 i 个评价等级的置信度,(D_i,β_i^k) 表示评价结果的置信分布形式;N 表示评价等级的个数;δ_i 表示第 i 个前提属性权重;θ_k 表示第 k 条规则的权重.

1.2 构建规则库

假设有 L 条数据,每条数据含有 T 个属性,生成 EBRB 规则库的步骤[30]:

步骤 1. 根据专家经验确定前提属性参考值个数 J_i 和评价属性参考值个数 N,确定具体的前提属性参考值$\{A_{ij}\}$,$i=1,2,\cdots,T$;$j=1,2,\cdots,J_i$ 和评价属性参考值$\{D_i\}$,$i=1,2,\cdots,N$.

步骤 2. 对步骤 1 得到的前提属性参考值和评价结果参考值,采用式(3)~(5),将数据集属性值和结果值转化为具有置信分布的形式:

$$S(x_i)=\{(A_{ij},\alpha_{ij})\}, \quad (2)$$

其中,$i=1,2,\cdots,T$,$j=1,2,\cdots,J_i$.假设数据 $X=\{x_i,i=1,2,\cdots,T\}$,规则第 i 个前提属性的第 j 个参考值为 $u(A_{ij})$,在同一属性下有$\{u(A_{ij})<u(A_{i(j+1)}),j=1,2,\cdots,J_i-1\}$.则 α_{ij} 计算为

$$\alpha_{ij}=\frac{u(A_{i(j+1)})-x_i}{u(A_{i(j+1)})-u(A_{ij})},$$
$$u(A_{ij})\leqslant x_i\leqslant u(A_{i(j+1)}),j=1,2,\cdots,J_i-1, \quad (3)$$
$$\alpha_{i(j+1)}=1-\alpha_{ij},u(A_{ij})\leqslant x_i\leqslant u(A_{i(j+1)}),$$
$$j=1,2,\cdots,J_i-1, \quad (4)$$
$$\alpha_{im}=0,m=1,2,\cdots j-1,j+2,\cdots,J_i-1. \quad (5)$$

同理可得规则评级结果的置信分布形式:

$$E(y)=\{(D_i,\beta_i)\},i=1,2,\cdots,N. \quad (6)$$

步骤 3. 将所有的输入数据通过步骤 2 将转化为如式(1)所示的置信规则形式,从而得到初始置信规则库.

步骤 4. 假设步骤 3 中选取规则 R_a 和规则 R_b,那么 R_a 和 R_b 间的 d_{SRA} 和 d_{SRC} 度量[20]:

$$d_{\mathrm{SRA}}(R_a,R_b)=\min_{i=1}^{T}\left(1-\sqrt{\sum_{j=1}^{J_i}(\alpha_{ij}^a-\alpha_{ij}^b)^2}\right), \quad (7)$$

$$d_{\mathrm{SRC}}(R_a,R_b)=1-\sqrt{\sum_{i=1}^{N}(\beta_{ij}^a-\beta_{ij}^b)^2}. \quad (8)$$

从而 R_a 和 R_b 的一致性度量为

$$Cons(R_a,R_b)=\exp\left(-\frac{(d_{\mathrm{SRA}}(R_a,R_b)/d_{\mathrm{SRC}}(R_a,R_b)-1.0)^2}{1/d_{\mathrm{SRA}}(R_a,R_b)^2}\right). \quad (9)$$

计算第 i 条规则与其他规则的不一致性:

$$Incons(i)=\sum_{j=1,j\neq i}^{L}(1-Cons(R_i,R_j)). \quad (10)$$

通过度量规则间的不一致程度,对每条规则权重进行调整,以此来降低由规则不一致性造成的影响:

$$\theta_k^*=\theta_k+\eta\times(1-Incons(k)/\xi), \quad (11)$$
$$\xi=\sum_{i=1}^{L}Incons(i), \quad (12)$$

其中,θ_k 是第 k 条规则的权重,η 是常数权重用来调节规则不一致性对规则权重的影响.若规则权重没有初始值,则规则权重可以表示为

$$\theta_k^*=1-Incons(k)/\xi. \quad (13)$$

1.3 推理机制

假设输入数据 $X=(x_1,x_2,\cdots,x_T)$,根据式(3)~(5)算出该输入数据的置信分布形式.通过该置信分布形式可计算该输入数据与第 k 条规则对于第 i 个属性的个体匹配度[31]:

$$d_i^k=\sqrt{\sum_{j=1}^{J_i}(\alpha_{ij}-\alpha_{ij}^k)^2}, \quad (14)$$
$$S_i^k=1-d_i^k, \quad (15)$$

其中,d_i^k 是关于输入数据与第 k 条规则的第 i 个属性的相似程度.因此,第 k 条规则的激活权重计算为

$$w_k=\frac{\theta_k\prod_{i=1}^{T}(S_i^k)^{\bar{\delta}_i}}{\sum_{k=1}^{L}\theta_k\prod_{i=1}^{T}(S_i^k)^{\bar{\delta}_i}},$$
$$\bar{\delta}_i=\frac{\delta_i}{\max\limits_{i=1,2,\cdots,T}\delta_i}, \quad (16)$$
$$\sum_{k=1}^{L}w_k=1. \quad (17)$$

其中,$0\leqslant w_k\leqslant 1(k=1,2,\cdots,L)$.

采用 ER 算法对激活的规则进行合成,获得带有置信分布的推理结果.ER 解析算法为[32]

$$\beta_n=\left\{\mu\left[\prod_{i=1}^{L}\left(w_i\beta_{n,i}+1-w_i\sum_{k=1}^{N}\beta_{k,i}\right)-\prod_{i=1}^{L}\left(1-w_k\sum_{k=1}^{N}\beta_{k,i}\right)\right]\right\}\Big/\left\{\left[1-\mu\left[\prod_{k=1}^{L}(1-w_k)\right]\right]\right\},\tag{18}$$

$$\mu=\left(\sum_{n=1}^{N}\prod_{i=1}^{L}\left(w_i+1-w_i\sum_{k=1}^{N}\beta_{k,i}\right)-(N-1)\prod_{i=1}^{L}\left(1-w_i\sum_{k=1}^{N}\beta_{k,i}\right)\right)^{-1}.\tag{19}$$

若求解问题为分类问题,则可以取最大置信度的评级结果等级作为输出:

$$f(x_i)=D_t,\quad t=\arg\max_{1\leqslant k\leqslant N}\{\beta_k\}.\tag{20}$$

若求解问题为回归问题,则可以通过计算期望效用值作为最后的输出:

$$f(x_i)=\sum_{k=1}^{N}\beta_k u(D_k).\tag{21}$$

2 基于析取范式的 EBRB

2.1 EBRB 中的数据缺失问题

Liu 等人[20]提出的 EBRB 系统能够通过历史数据生成置信规则库,避免了反复迭代的参数优化过程,但对于含有缺失数据的数据集,原始的 EBRB 方法缺少相应的处理措施,必须要对数据集进行预处理,否则无法直接生成规则库.例如输入数据为 $X=(x_1,x_2,\cdots,x_{t-1},\text{none},x_{t+1},\cdots,x_T)$,第 t 个属性为数据缺失,则式(3)~(5)无法计算得到该属性的置信分布形式,所以该条数据无法直接生成规则.另一方面,因为 EBRB 系统采用合取范式连接属性,含有缺失项的数据就无法对规则进行激活.

Chang 等人[27]提出的 DBRB 系统通过改变属性连接方式,能够有效地避免组合爆炸问题.同时析取范式的规则形式面对含有缺失项的数据具有一定的优势,因为当规则中有 1 个属性的匹配度不为 0 时,该条规则的激活权重便不为 0,所以当数据含有缺失值时,仍可以计算其他属性值对规则的激活贡献.例如:上述的输入数据 x 中 x_t 为缺失项,当该数据去生成规则时,虽然无法知道缺失值的置信分布,但仍可以利用未缺失的数据 $x_{i,i\neq t}$ 去计算该条规则部分属性的匹配度.虽然不可避免地会对 BRB 系统的推理准确性产生影响,但 DBRB 系统仍可以进行推理.由于数据缺失了部分信息,该数据生成的规则具有更大的不确定性,容易在 ER 合成时产生误导作用,所以应该降低该规则的权重以减少该条规则对推理结果的影响.

综上所述,本文方法结合前面所提 2 者的特点,建立一个基于析取范式的扩展置信规则库.该方法主要通过不完整数据生成含有不完整规则的扩展置信规则库,并引入衰减因子对不完整规则权重进行调节.该方法能够更充分地利用已有的信息,同时在不增加系统复杂性的情况下,保证推理的准确性.

2.2 规则表示

利用析取范式规则和 EBRB 快速构建规则库的优点,建立基于析取范式的 EBRB 系统,规则表示:

$$\begin{aligned}R_k:&\text{IF } U_1 \text{ is } \{(A_{11},\alpha_{11}^k),(A_{12},\alpha_{12}^k),\cdots,(A_{1J_1},\alpha_{1J_1}^k)\}\vee U_2 \text{ is } \{(A_{21},\alpha_{21}^k),(A_{22},\alpha_{22}^k),\cdots,(A_{2J_2},\alpha_{2J_2}^k)\}\vee\cdots\vee\\&U_T \text{ is } \{(A_{T1},\alpha_{T1}^k),(A_{T2},\alpha_{T2}^k),\cdots,(A_{TJ_T},\alpha_{TJ_T}^k)\}\\&\text{THEN } \{(D_1,\beta_1^k),(D_2,\beta_2^k),\cdots,(D_N,\beta_N^k)\}\end{aligned}$$

其中,θ_k 表示规则权重,$\delta_1,\delta_2,\cdots,\delta_T$ 表示属性权重,$\varepsilon_1,\varepsilon_2,\cdots,\varepsilon_k$ 表示衰减因子;

$$\beta\geqslant 0,\alpha_{ij}^k\geqslant 0,\sum_{j=1}^{N}\beta_j^k\leqslant 1,\sum_{j=1}^{J_i}\alpha_{ij}^k\leqslant 1,\tag{22}$$

其中,ε_i 表示由不完整数据生成的第 i 条规则的衰减因子,其余参数含义与式(1)一致.

2.3 新的激活权重计算公式

由于规则表示形式改变,属性间连接具有逻辑或的关系,所以式(15)中的连乘符号不再适用.新的激活权重定义为

$$w_k=\frac{\xi_k\sum_{i=1}^{T}S_i^k\bar{\delta}_i}{\sum_{k=1}^{L}\left[\sum_{i=1}^{T}S_i^k\bar{\delta}_i\right]},\tag{23}$$

$$\xi_k=\theta_k\varepsilon_k,\bar{\delta}_i=\frac{\delta_i}{\max\limits_{i=1,2,\cdots,T}\delta_i}.\tag{24}$$

2.4 前提属性参考值个数的可变性

在 DBRB 中前提属性参考值的个数通常由领域专家经验来设置,同时前提属性参考值个数将决定规则库的规则数量,而在 Liu 等人[20]提出的 EBRB 中除了通过领域专家经验来获取,还可以通过对数据集的属性值所在区间中均匀取 N 个值来获取前属性参考值.对于析取范式的 EBRB,当前提属性参考值取值个数进行改变时,系统的推理精度会受到影响,以函数拟合实验[30]为例,非线性函数表达式:

$$f(x)=x\ \sin(x^2),0\leqslant x\leqslant 3.\tag{25}$$

选取 5 个值−2.5,−1,1,2,3 作为评价结果等级效用值,比较前提属性 x 的参考值个数 N 在取值不同时的推理性能.在区间[0,3]中均匀选取 500 个点作训练集,再均匀选取 1 000 个点作为测试集.图 1 为训练集,图 2 为系统在不同参考值个数上的拟合结果:

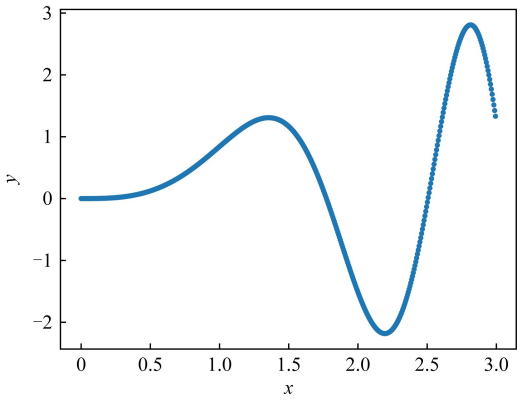

Fig. 1 Data of function fitting

图 1 函数拟合数据集

从图 2 中可知当 N 取值较低时,函数拟合效果不理想,但随着取值个数的增加,拟合效果有明显的改善,但取值个数到达 30 后拟合精度没有明确的增加.从表 1 可知随着参考值个数的增加,推理时间逐渐增多,系统性能消耗逐渐增大.

Table 1 Function Fitting Effect of Different N

表 1 不同 N 值下的函数拟合效果

N	规则数	平均激活规则数	平均推理时间/ms	平均绝对误差
5	500	227	5.899	0.450 199
10	500	110	6.598	0.182 863
15	500	72	7.995	0.105 506
30	500	35	12.643	0.056 142
50	500	21	19.145	0.046 686
70	500	15	24.724	0.044 714

由表 1 可知,参考值个数取值的不同会影响构建后规则库中规则的多样性,参考值个数取低数量时,相同置信分布结构的规则前件将增多,对于析取范式规则来说,属性更容易被匹配,从而容易造成更多的激活规则;当参考值个数增多,生成的规则差异性增大,不容易激活多条规则,但由于参考值个数增多,系统在计算个体匹配度时要花费更多的开销.因此,对前提属性参考值个数的合理选择能够使系统的推理性能更优.

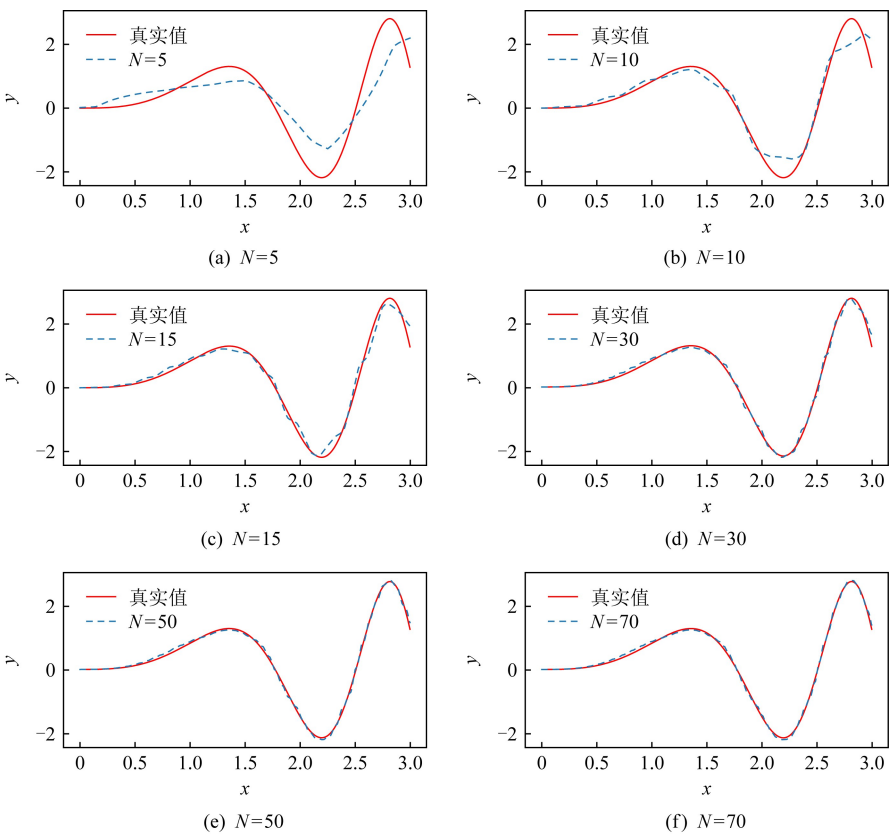

Fig. 2 Function fitting effect of different N

图 2 在不同 N 下的函数拟合效果

2.5 对缺失数据的处理

析取范式的 EBRB 对缺失数据处理的核心思想是:对不完整数据生成不完整规则,引入衰减因子,根据缺失值的属性权重调节规则权重.假设数据 $X=\{x_1,x_2,\cdots,x_T\}$,属性权重 $\delta=\{\delta_1,\delta_2,\cdots,\delta_T\}$.当数据第 t 个属性值缺失数据时,构建基于析取范式的规则将对 x_t 的置信分布情况完全无知,此时该规则处于不完整状态.根据缺失属性权重值,设置规则权重的衰减因子:

$$\varepsilon_i=\frac{\sum_{i=1}^{T}\delta_i-\sum_{j=t_1}^{t_n}\delta_j}{\sum_{j=1}^{T}\delta_i},\tag{26}$$

其中,$\delta_j(j=t_1,t_2,\cdots,t_n)$为缺失数据的属性权重.可知,完整数据生成的规则对应的衰减因子为 1,不完整数据生成规则对应的衰减因子是由缺失属性的权重和个数所决定的.具体步骤为:

步骤 1. 确定前提属性参考值个数,确定前提属性参考值和评价结果参考值;

步骤 2. 对没有缺失属性的数据使用式(3)~(5)计算个体匹配度;

步骤 3. 当属性值缺失时,则该属性的置信分布是完全无知的,所以设置该属性的个体匹配度为 0,即表示对于该缺失的属性值无法匹配任意一个属性候选值;

步骤 4. 使用式(3)~(5)计算评价结果的属性置信分布,使用式(6)~(9)计算该条规则的不一致性程度;

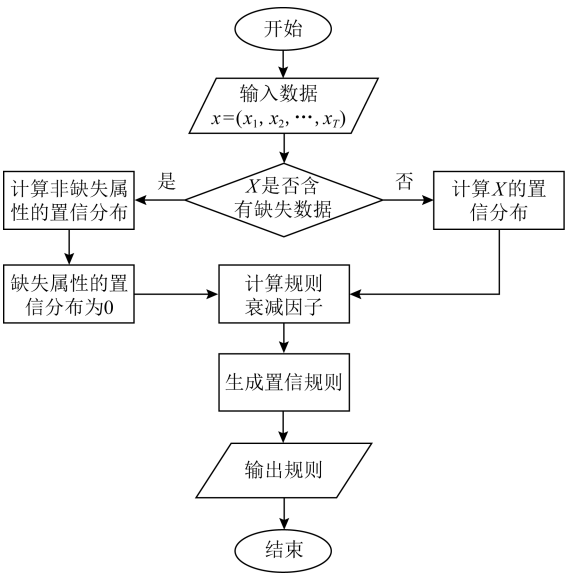

Fig. 3 Rule generation flowchart

图 3 规则生成流程图

步骤 5. 使用式(26)计算规则衰减因子;

步骤 6. 输出带有衰减因子的不完整规则.

析取范式 EBRB 生成规则流程如图 3 所示.

3 实验分析

本节实验选取公用测试数据库中的 5 个数据集 iris, wine, seeds, transfusion, mammographic 进行测试效果,表 2 提供给数据集详细信息.首先寻找本文方法在各个数据集中的最优前提属性参考值个数的选择,其次在最优参考值个数下,将本文方法与其他 EBRB 系统进行对比,验证本文方法的有效性,最后测试在不同程度的数据缺失率下本文的表现,并与传统的填补方法作对比.实验环境为:Intel Corei5@1.4 GHz;16 GB 内存;macOS Catalina 操作系统;编程环境为 python3.8.

Table 2 Details of Datasets

表 2 数据集说明

序号	名称	数量	属性数	分类数
1	iris	150	4	3
2	seeds	210	7	3
3	wine	178	13	3
4	transfusion	748	4	5
5	mammographic	830	5	2

3.1 与其他 EBRB 方法进行对比

3.1.1 确定前提属性参考值的个数

从 2.4 节可知,析取范式 EBRB 的前提参考属性的不同个数会对最后推理结果产生很大的影响.因此需要对前提属性参考值的取值个数进行比较,并采用分类准确率和推理时间作为评价依据.

在构建析取范式的 EBRB 时,分别对前提属性参考值个数 N 取值为 2,4,5,10,15,20 个,在属性值的最大最小值构成的区间中均匀地取 N 个点作为前提属性参考值的效用值.

当对 N 取不同值时,实验结果如图 4 所示.可知不同数据集对 N 取值相同时有不同的表现情况.对于 iris 数据集,当属性参考值为 5~10 时,分类准确率能达到 96%,而属性参考值个数超过 10 时,分类精度开始下降,同时推理时间增加,因此 iris 数据集的属性参考值个数的最优选择是 5;seeds 数据在属性参考值个数取到 4 时,能达到较好的准确率和较低的推理时间;在 transfusion 数据集中,当参考值个数取 2 时,能使系统的推理性能最优;mammo-

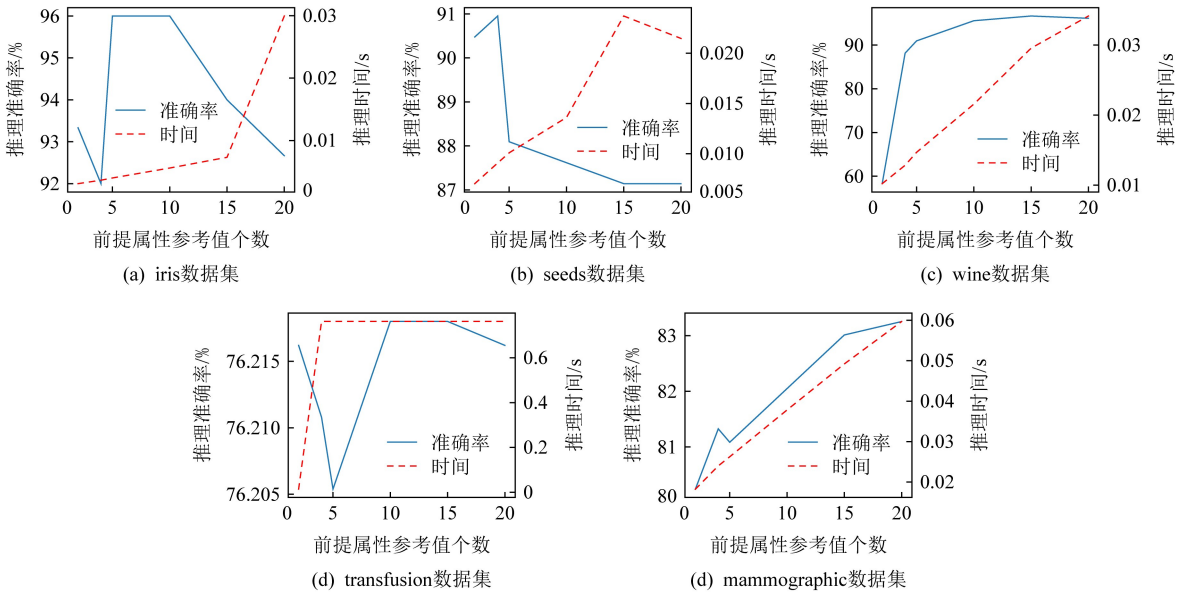

Fig. 4 Classification accuracy and time consumption of different number of reference in datasets

图 4 5 个数据集中不同参考值个数的准确率和推理时间

graphic 数据集会随着参考值个数的增加而显著增加,当个数超过 15 时,准确率增加变缓,但推理时间仍呈线性关系增加,因此参考值个数选择 15 时为最优值;wine 数据集在参考值个数取到 10 时达到较高的准确率,同时保证了较低的系统推理时间.

3.1.2 对比实验

为了验证析取范式 EBRB 在完整数据集下的有效性,下面将与 Liu-EBRB[20],DRA-EBRB[33],CABRA-EBRB[34],EBRB-C[35],VP-EBRB[36],MVP-EBRB[36]这些 EBRB 优化模型进行对比实验.在对比实验中,本文方法的前提属性参考值个数采用最优推理效率时的个数,其他 EBRB 算法均默认取 5 个,同时采用 10 折交叉验证方法进行分类准确率的比较.

从表 3 可知,本文方法在 mammographic 数据集上表现明显优于其他 6 种 EBRB,并在 wine 和 iris 数据集上能够和 CABRA-EBRB 具有相同的推理性能,明显优于其他 EBRB 方法.虽然在其他 2 个数据集上的表现一般,但准确率和 Liu 等人[20]提出的原始 EBRB 方法效果基本相持平.

该实验表明:本文方法在完整数据集上与其他 EBRB 模型相比有不错的表现.其中 mammographic 数据集上表现最好,在 iris 和 wine 数据集上表现位列第 2,在 transfusion 和 seeds 数据集上表现一般,但仍能保持较好的准确率.下面对含有缺失数据的问题进行验证.

Table 3 Compare with Other EBRB Accuracy Rates

表 3 与其他 EBRB 准确率进行对比 %

方法	iris	seeds	wine	transfusion	mammographic
Liu-EBRB	95.33	91.43	96.24	76.14	79.70
DRA-EBRB	95.50	92.02	96.46	76.57	78.39
VP-EBRB	95.13	92.57		77.33	
MVP-EBRB	95.87	92.38		80.36	
CABRA-EBRB	96.00	92.38	96.63	72.07	79.52
EBRB-C	95.73	93.24	96.32	76.88	77.64
本文方法	96.00	90.95	96.63	76.22	83.01

3.2 缺失数据的对比

本节实验进行在不同的数据缺失率和缺失程度下进行对比.实验中,数据集采用 2 种缺失模式:单属性随机缺失和多属性随机缺失,数据集中数据缺失率设置为 10%~90%.在单属性随机缺失模式中,

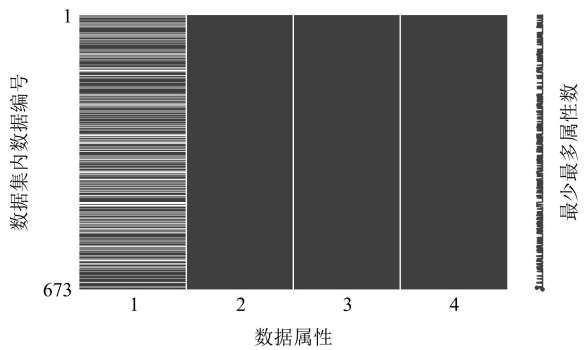

Fig. 5 The first attribute randomly miss 30% of data

图 5 第 1 个属性随机缺失 30%的数据

每次在一个属性中设置缺失项,遍历每个属性,对最后结果取均值,如图 5 示例,数据在第 1 个属性上有 30%的缺失率.在多属性随机缺失模式中,缺失数据位置均为随机产生,缺失率为缺失数据项占全部数据项的百分比,如图 6 示例,数据存在 10%的缺失,且缺失可能产生在任何位置,该示例中,数据集中既有缺失 3 个属性值的数据,又有不存在缺失的数据.

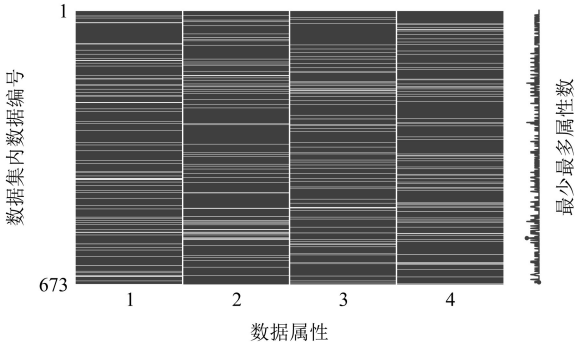

Fig. 6 Multi attribute randomly miss 10% of data

图 6 多属性随机缺失 10%数据

3.2.1 与 EBRB 对比实验

在 3.1.2 节的 5 个数据集中本文方法的表现性能与 Liu-EBRB 方法相近,故选择该 EBRB 进行对比实验,并采用平均值、中位数和众数对缺失项进行填补,使 Liu-EBRB 能正常进行推理,为了比较方法的推理性能,对于零激活的输入,默认为分类错误.本次实验采用 20 次 10 折交叉验证取均值的方式进行验证.

从图 7 可知,在单属性缺失模式下,本文方法对不同缺失率的数据集具有更平稳的表现,在 wine 和 mammographic 数据集中,本文方法明显优于统计方法,并且在随着数据的缺失率逐渐上升,本文方法的分类准确率下降平缓;在 iris 数据集中,在较低的缺失率下,本文方法略低于统计方法,但随着缺失率逐渐上升,传统方法开始逐渐下滑,当缺失率超过 50%时,本文方法的准确率明显高于统计方法;在 transfusion 和 seeds 数据集中,本文方法在较低缺失率的数据集下的表现比统计方法来得差,但面对不同的缺失率,本文方法具有更平稳的表现,随着缺失率增高,采用统计方法填补更容易造成零激活问题,因为填补数据单一,使得单个属性值的多样性变少,容易使缺失数据属性的个匹配度为 0,从而造成零激活.而本文方法尽管在缺失数据属性的个体匹配度为 0,但规则可以被其他属性的匹配度所激活,因此本文方法在高缺失率的数据集中仍能有较好的表现.

从图 8 可知,在多属性随机缺失模式下,本文方法具有更好的推理性能.在 iris 和 seeds 数据中,在

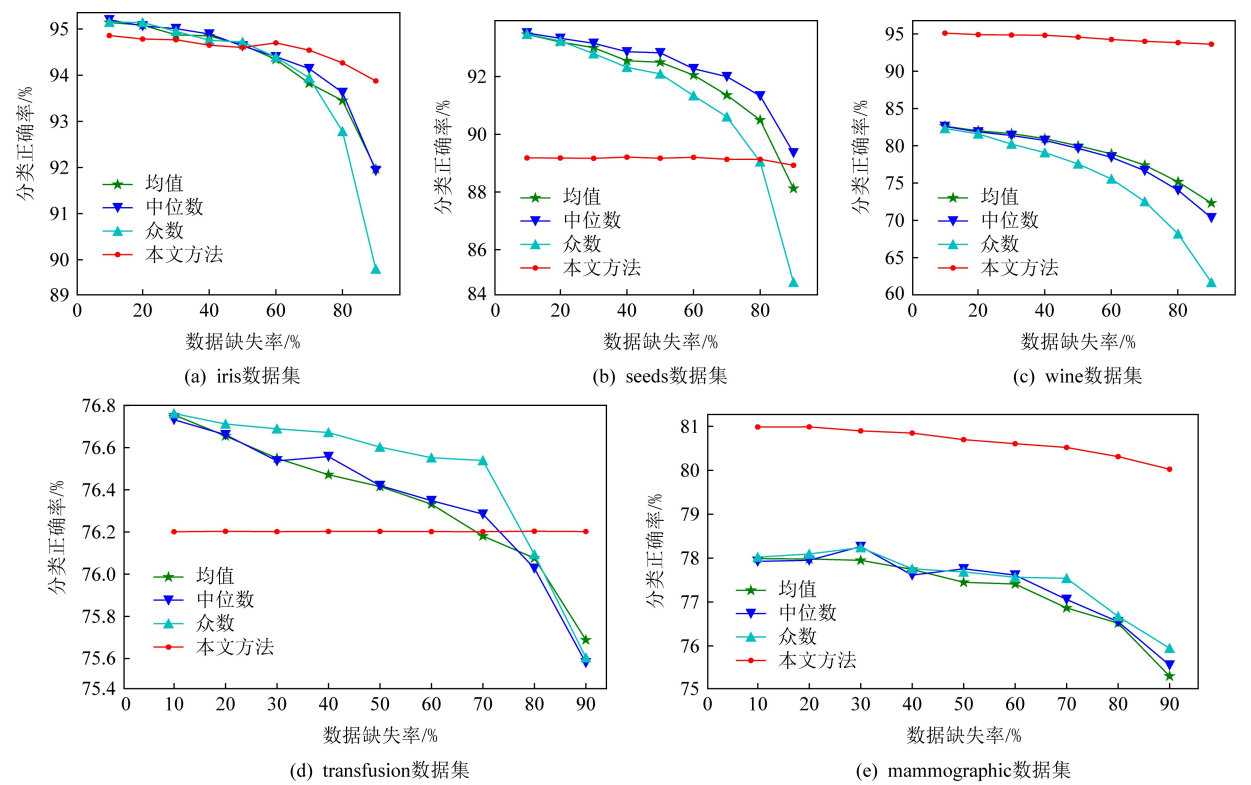

Fig. 7 Experiments of EBRB single attribute randomly miss

图 7 EBRB 单属性随机缺失实验

低缺失率情况下,本文方法能和统计方法保持相同的准确率,并且在高缺失率下能具有更高的准确率;在 wine 数据集中,Liu-EBRB 表现性能较坏,因为由于采用统计方法进行填补缺失值,使得构建规则库的数据多样性减少,在高缺失率下,很容易大量产生零激活,从而造成较低的分类准确率;在 mammographic 数据集中,本方法能在数据集缺失较严重的情况下,仍保持较高的精度,而统计方法明显会随着缺失数据地增多而减少准确率;在 transfusion 数据集中,在低缺失率情况下,本文方法能保证较好的准确率,同时当缺失率大于 50%时,本文方法能明显优于统计方法.因为当数据缺失率较高的情况下,容易出现属性值全部缺失的情况,此时该数据生成规则的衰减因子 $\varepsilon=0$,根据式(23)可知该规则实际上不参与最后的 ER 推理,本文方法便等价于删去了属性值全部缺失的数据,从而生成更小规模的规则库,并使部分缺失的数据仍能为 ER 推理贡献一部分置信度.

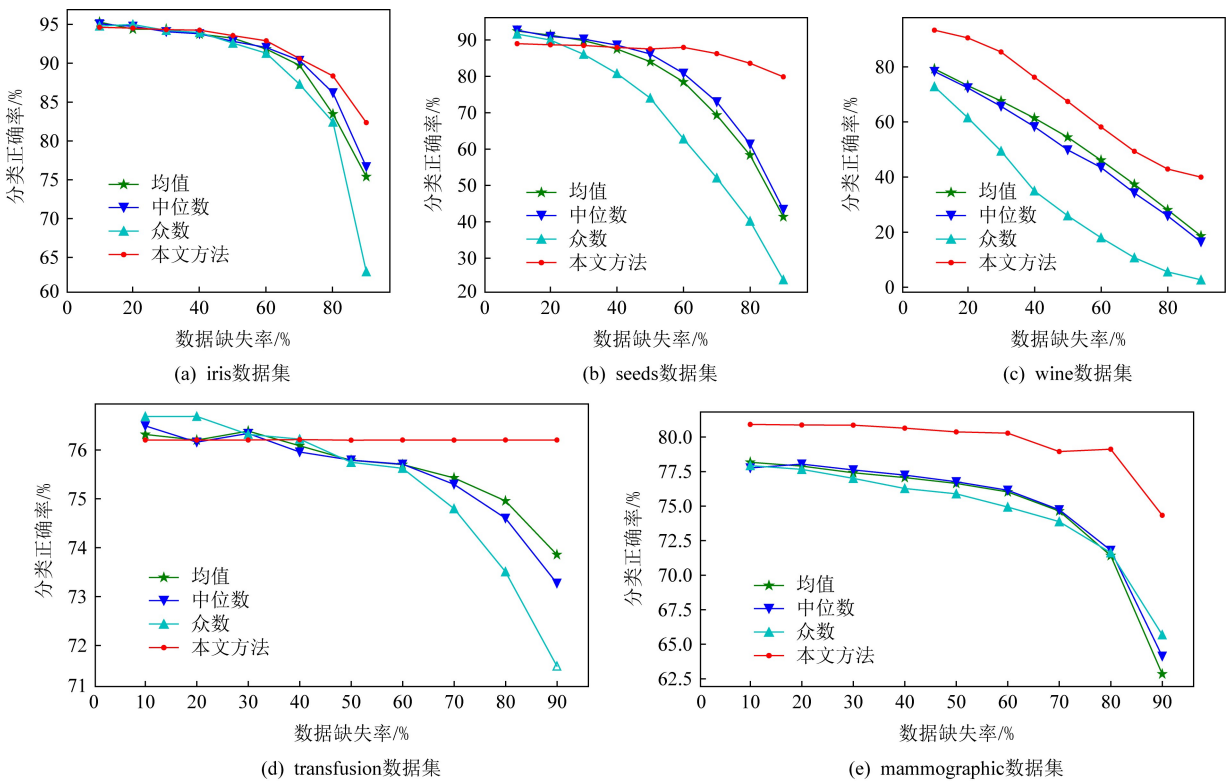

Fig. 8 Experiments of EBRB multi attribute randomly miss

图 8 EBRB 多属性随机缺失实验

3.2.2 与 DBRB 对比实验

根据 2.1 节可知,DBRB 能够对含有缺失项的数据进行推理,因此本节实验主要对比方法与 DBRB 方法在不同缺失率下的推理准确率.在 DBRB 方法中,为了不影响 DBRB 的推理过程,将缺失数据对应的个体匹配度设置为 0,即缺失项在计算规则激活程度时不提供任何作用.

图 9 为单属性随机缺失模式下不同缺失率的对比实验结果.从图 9 中可知本文方法均高于 DBRB 方法.图 9 中本文方法的曲线下降更加平缓,而 DBRB 推理准确率会随着数据缺失率的增大而下降,原因在于随着数据缺失量的增大,缺失数据的属性无法充分地进行训练,容易得到对含有缺失项的属性权重设置为 0 的训练结果,当该属性为数据集的重要特征时很容易在推理过程中产生误判,造成不准确的推理结果.而本文方法借鉴 EBRB 构造规则库的方法,跳过对规则参数设置的步骤,直接利用数据生成规则及参数,从而避免了参数训练造成的问题,因此本文方法在单属性缺失模式下有更好的表现.

图 10 为在多属性随机缺失模式下不同缺失率下的对比实验结果.在 iris,seeds,transfusion 数据集上,本文方法明显优于 DBRB 方法,原因为在多属性缺失模式下,同样存在 DBRB 难以训练参数的情况,同时对于含有缺失项的数据很难充分利用数据中剩余的信息,因此 DBRB 推理准确率仍会随着数据缺失程度的增大而减少,而本文方法采用数据

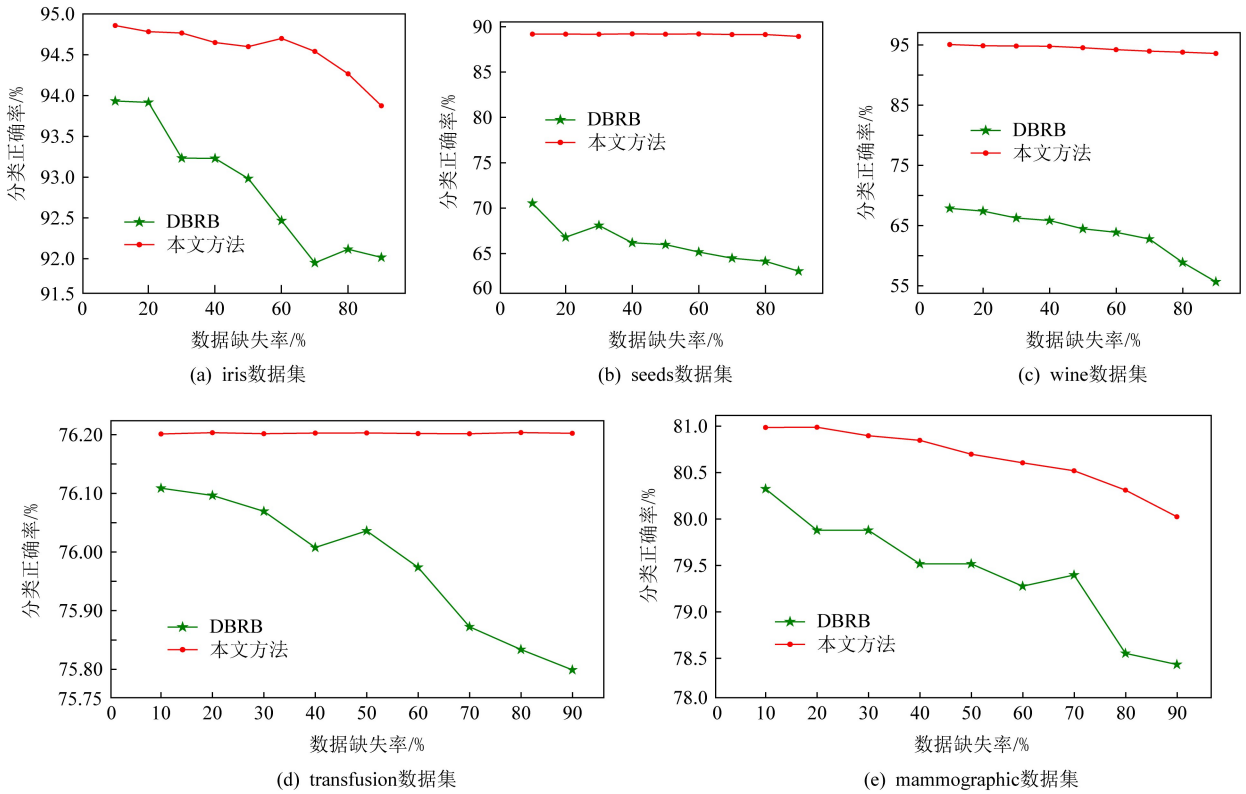

Fig. 9 Experiments of DBRB single attribute randomly miss

图 9 DBRB 单属性随机缺失实验

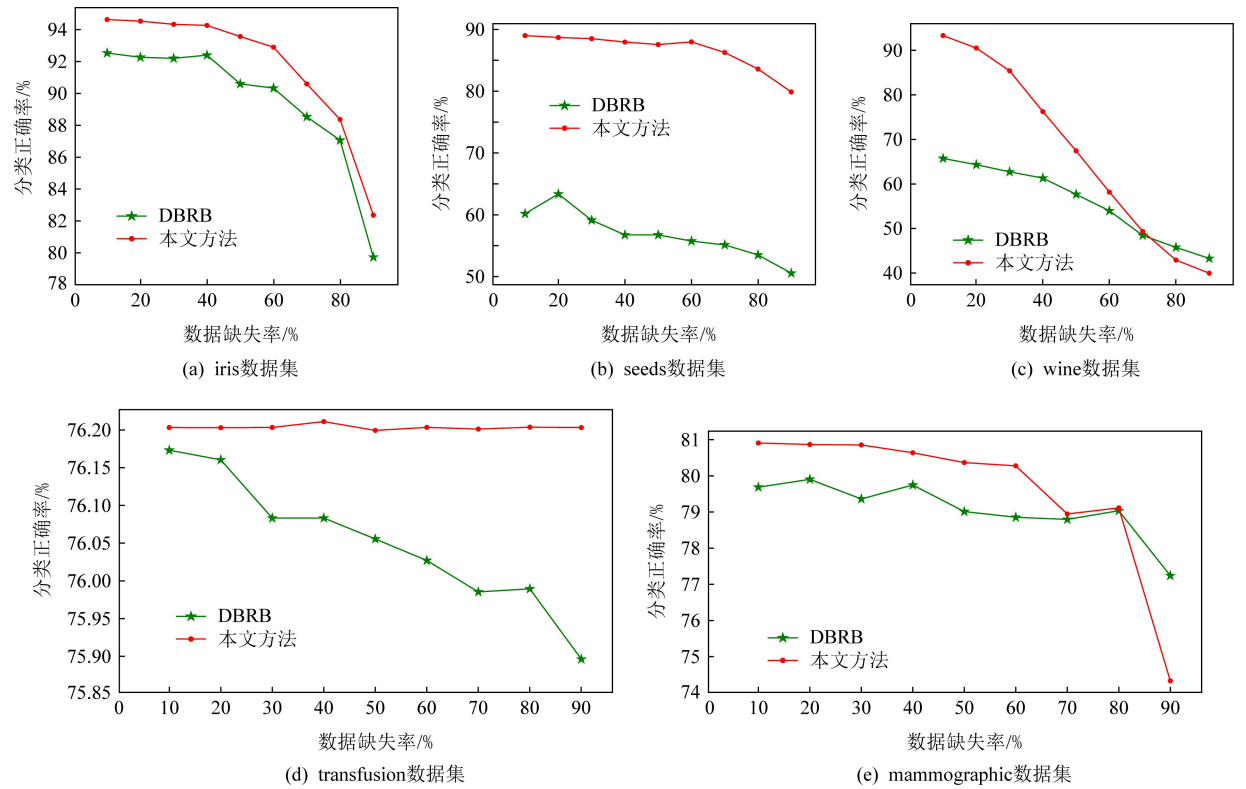

Fig. 10 Experiments of DBRB multi attribute randomly miss

图 10 DBRB 多属性随机缺失实验

生成规则的方法构建规则库,当数据缺失程度较大时,构建的规则库中会存在大量不完整规则,这些规则通过衰减因子调节规则权重,仍将参与最终的ER合成,这使得本文方法能够更有效的利用数据中的信息.但当缺失程度过大时,规则库中存在大量不完整规则,造成规则推理结果的不确定性增加,使得最后的推理准确率出现下降,在iris,wine,mammographic数据集上可知,当数据缺失超过70%时,本文方法的推理准确率开始快速下滑.

本节实验结果表明:适当的前提属性参考值个数能够对本文方法的推理性能产生明显影响,本文方法在完整数据集中能够保证一定的准确度.同时本文方法在不同缺失率和缺失模式下的数据集中能保证推理准确率的稳定性,在与EBRB方法对比实验中,低缺失率的数据集能得到较好的准确率,在高缺失率的数据集中本文方法能明显优于通过统计填充方法,在与DBRB方法对比实验中,本文方法能够有更好的准确率,并且避免了参数训练带来的性能消耗.上述实验证明了本文方法面对含有缺失的数据集是有效的.

4 结束语

本文针对BRB系统中数据缺失问题,提出了析取范式的扩展置信规则库解决方法.该方法通过不完整数据集生成含有不完整规则的析取范式规则库,并使用衰减因子调节不完整规则权重,使不完整规则更合理地参与到系统推理过程中.为了验证新推理方法的有效性,本文选择了mammographic,iris,wind,seeds,transfusion数据集进行验证,并选择Liu-EBRB和DBRB进行不同数据缺失率和缺失模式下的对比实验,取得了预期效果.

作者贡献声明:刘永裕完成实验和并撰写论文;巩晓婷协助撰写论文;方炜杰协助论文实验;傅仰耿提出指导意见并修改论文.

参 考 文 献

[1] Rahman M G, Islam M Z. KDMI: A novel method for missing values imputation using two levels of horizontal partitioning in a data set [C/OL] //Proc of the 9th Int Conf on Advanced Data Mining and Applications (ADMA 2013). Berlin: Springer, 2013 [2020-07-09]. https://xueshu.baidu.com/usercenter/paper/show?paperid=1af936b1b581db9cd20beeb61ef3ce09&site=xueshu_se&hitarticle=1

[2] Gustavo E A P A, Monard M C. An analysis of four missing data treatment methods for supervised learning [J]. Applied Artificial Intelligence, 2003, 17(5/6): 519-533

[3] Liu Zhunga, Liu Yong, Dezert J. Classification of incomplete data based on belief functions and k-nearest neighbors [J]. Knowledge-Based Systems, 2015, 89: 113-125

[4] Schneider T. Analysis of incomplete climate data: Estimate of mean values and covariance matrices and imputation of missing values [J]. Journal of Climate, 2001, 14(5): 853-871

[5] Dempster A P. Maximum likelihood from incomplete data via the EM algorithm [J]. Journal of the Royal Statistical Society: Series B Methodological, 1977, 39(1): 1-38

[6] Ichihashi H, Honda K, Notsu A, et al. Fuzzy c-means classifier with deterministic initialization and missing value imputation [C] //Proc of 2007 IEEE Symp on Foundations of Computational Intelligence. Piscataway, NJ: IEEE, 2007: 214-221

[7] Chechik G, Heitz G, Elidan G, et al. Max-margin classification of incomplete data [C] //Proc of Conf on Advances in Neural Information Proccessing Systems. Cambridge, MA: MIT Press, 2006: 233-240

[8] Wang Shuangcheng, Yuan Senmiao. Research on learning Bayesian networks structure with missing data [J]. Journal of Software, 2004, 15(7): 1042-1048 (in Chinese)
(王双成,苑森淼. 具有丢失数据的贝叶斯网络结构学习研究[J]. 软件学报, 2004, 15(7): 1042-1048)

[9] Chen Jingnian, Huang Houkuan, Tian Fengzhan, et al. Selective Bayes classifiers for incomplete data [J]. Journal of Computer Research and Development, 2007, 44(8): 1324-1330 (in Chinese)
(陈景年,黄厚宽,田凤占,等. 用于不完整数据的选择性贝叶斯分类器[J]. 计算机研究与发展, 2007, 44(8): 1324-1330)

[10] Yang Jianbo, Liu Jun, Wang Jin. Belief rule-base inference methodology using the evidential reasoning approach-RIMER [J]. IEEE Transactions on Systems, Man, and Cybernetics: Part A Systems and Humans, 2006, 36(2): 266-285

[11] Wang Yingming, Yang Jianbo, Xu Dongling. Consumer preference prediction by using a hybrid evidential reasoning and belief rule-based methodology [J]. Expert Systems with Applications, 2009, 36(4): 8421-8430

[12] Yang Ying, Fu Chao, Chen Yuwang. A belief rule based expert system for predicting consumer preference in new product development [J]. Knowledge-Based Systems, 2016, 94: 105-113

[13] Yang Jianbo, Wang Yingming, Xu Dongling. Belief rule-based methodology for mapping consumer preferences and setting product targets [J]. Expert Systems with Applications, 2012, 39(5): 4749-4759

[14] Li Gailing, Zhou Zhijie, Hu Changhua. A new safety assessment model for complex system based on the conditional generalized minimum variance and the belief rule base [J]. Safety Science, 2017, 93: 108-120

[15] Sun Chuan, Wu Chaozhong, Chu Duanfeng. A recognition model of driving risk based on belief rule-base methodology [J/OL]. International Journal of Pattern Recognition & Artificial Intelligence, 2017 [2020-07-09]. https://www.worldscientific.com/doi/10.1142/S0218001418500374

[16] Li Gailing, Zhou Zhijie, Hu Changhua. An optimal safety assessment model for complex systems considering correlation and redundancy [J]. International Journal of Approximate Reasoning, 2019, 104: 38-56

[17] Kong Guilan, Xu Dongling, Liu Xiaobo. Applying a belief rule-base inference methodology to a guideline-based clinical decision support system [J]. Expert Systems, 2019, 26(5): 391-408

[18] Zhou Zhigao, Liu Fang, Li Lingling. A cooperative belief rule based decision support system for lymph node metastasis diagnosis in gastric cancer [J]. Knowledge-Based Systems, 2015, 85: 62-70

[19] Rahaman S. Diabetes diagnosis expert system by using belief rule base with evidential reasoning [C] //Proc of Int Conf on Electrical Engineering & Information Communication Technology. Piscataway, NJ: IEEE, 2015: 1-5

[20] Liu Jun, Martinez L, Calzada A. A novel belief rule base representation, generation and its inference methodology [J]. Knowledge-Based Systems, 2013, 53: 129-141

[21] Calzada A, Liu Jun, Wang Hui. Application of a spatial intelligent decision system on self-rated health status estimation [J]. Journal of Medical Systems, 2015, 39(11): 138-138

[22] Wang Yingming, Ye Feifei, Yang Longhao. Extended belief rule based system with joint learning for environmental governance cost prediction [J/OL]. Ecological Indicators, 2020, 111: 106070 [2020-07-09]. https://www.sciencedirect.com/science/article/pii/S1470160X20300078

[23] Espinilla M, Medina J, Calzada A, et al. Optimizing the configuration of an heterogeneous architecture of sensors for activity recognition, using the extended belief rule-based inference methodology [J]. Microprocessors and Microsystems, 2016, 52(Jul): 381-390

[24] Yang Longhao, Liu Jun, Wang Yingming, et al. Comparative analysis on extended belief rule-based system for activity recognition [C] //Proc of Data Science and Knowledge Engineering for Sensing Decision Support. Singapore: World Scientific, 2018: 430-436

[25] Yu Meng, Huang Jian, Kong Jiangtao. Belief rule-base inference methodology with incomplete input [J]. Journal of Harbin Institute of Technology, 2019, 51(4): 51-59 (in Chinese)
(鱼蒙, 黄健, 孔江涛. 输入信息不完整的置信规则库推理方法[J]. 哈尔滨工业大学学报, 2019, 51(4): 51-59)

[26] Wang Xiaoyan, Sun Jianbin, Zhao Qingsong, et al. Belief rule-based approach for capability satisfactory evaluation with incomplete information [J]. Systems Engineering and Electronic, 2019, 41(11): 2507-2513 (in Chinese)
(王小燕, 孙建彬, 赵青松, 等. 不完备信息条件下基于置信规则库的能力满足度评估方法[J]. 系统工程与电子技术, 2019, 41(11): 2507-2513)

[27] Chang Leilei, Jiang jiang, Sun Jianbin. Disjunctive belief rule base spreading for threat level assessment with heterogeneous, insufficient, and missing information [J]. Information Sciences, 2019, 416: 106-131

[28] Li Shaohua, Feng Jingying, He Wei. Health assessment for a sensor network with data loss based on belief rule base [J]. IEEE Access. 2020, 8: 126347-126357

[29] Liang L R, Looney C G, Mandal V. Fuzzy-inferenced decisionmaking under uncertainty and incompleteness [J]. Applied Soft Computing, 2011, 11(4): 3534-3545

[30] Chen Nannan, Gong Xiaoting, Fu Yanggeng. Extended belief rule base method based on improved rule activation rate [J]. CAAI Transaction on Intelligent Systems, 2019, 14(6): 1179-1188 (in Chinese)
(陈楠楠, 巩晓婷, 傅仰耿. 基于改进规则激活率的扩展置信规则库推理方法[J]. 智能系统学报, 2019, 14(6): 1179-1188)

[31] Fu Yanggeng, Zhuang Jinhui, Chen Yupeng, et al. A framework for optimizing extended belief rule base systems with improved Ball trees [J/OL]. Knowledge-Based Systems, 2020, 210: 106484 [2020-07-09]. https://www.sciencedirect.com/science/article/pii/S0950705120306134

[32] Wang Yingming, Yang Jianbo, Xu Dongling. Environmental impact assessment using the evidential reasoning approach [J]. European Journal of Operational Research, 2006, 174(3): 1885-1913

[33] Calzada A, Liu Jun, Wang Hui. A new dynamic rule activation method for extended belief rule-based systems [J]. IEEE Transactions Knowledge & Data Engineering, 2015, 27(4): 880-894

[34] Yang Longhao, Wang Yingming, Fu Yanggeng. A consistency analysis-based rule activation method for extended belief-rule-based systems [J]. Information Sciences, 2018, 445: 50-65

[35] Yang Longhao, Liu Jun, Wang Yingming. New activation weight calculation and parameter optimization for extended belief rule-based system based on sensitivity analysis [J]. Knowledge and Information Systems, 2019, 60: 837-878

[36] Lin Yanqing, Fu Yanggeng, Su Qun. A rule activation method for extended belief rule base with VP-tree and MVP-tree[J]. Journal of Intelligent & Fuzzy Systems, 2017, 33(6): 3695-3705

Liu Yongyu, born in 1995. Master. His main research include intelligent decision technology and belief rule base inference, etc.

刘永裕,1995 年生.硕士.主要研究方向为智能决策技术、置信规则库推理等.

Gong Xiaoting, born in 1982. PhD candidate, lecturer. Her main research interests include multi-criteria decision making under uncertainty and information hiding technology, etc.

巩晓婷,1982 年生.博士研究生,讲师.主要研究方向为不确定多准则决策、信息隐藏技术等.

Fang Weijie, born in 1995. PhD candidate. His main research interests include multi-objective optimization, intelligent decision technology, rule-based inference, big data analysis and data mining, etc.

方炜杰,1995 年生.博士研究生.主要研究方向为多目标优化、智能决策技术、规则推理方法、大数据分析和数据挖掘等.

Fu Yanggeng, born in 1981. PhD, professor, PhD supervisor, senior member of CCF. His main research interests include data mining, machine learning, and intelligent decision support systems, etc.

傅仰耿,1981 年生.博士,教授,博士生导师,CCF 高级会员.主要研究方向为数据挖掘与机器学习、智能决策支持系统等.