实体异构性下证据链融合推理的多属性群决策

沈 江 [1] 余海燕 [1] 徐 曼 [2]

摘 要 针对多属性群决策中可解释性证据融合推理的实体异构性问题, 给出了一个实体异构性下证据链融合推理的多属性群决策方法. 基于证据推理理论, 引入证据链关联的概念, 从多数据表提供的数据矩阵中获取可区分的近邻证据集, 推导了各数据表的相似度矩阵, 并构建半正定矩阵的二次优化模型, 共享群决策专家的经验知识. 使用 Dempster 正交规则, 论证了异构实体之间可解释性推理中可信度融合的合理性, 并使用证据融合规则集成各个数据表的近邻证据中获得的可信度, 验证了调和多源异构数据中不一致信息的有效性. 通过具有实体异构性的心脏病多决策数据诊断实例说明了方法的可行性与合理性.

关键词 实体异构性, 证据链关联, 相似度矩阵, 融合推理, 群体智慧

引用格式 沈江, 余海燕, 徐曼. 实体异构性下证据链融合推理的多属性群决策. 自动化学报, 2015, **41**(4): 832−842

DOI 10.16383/j.aas.2015.c140650

Heterogeneous Evidence Chains Based Fusion Reasoning for Multi-attribute Group Decision Making

SHEN Jiang[1] YU Hai-Yan[1] XU Man[2]

Abstract In multi-attribute group decision making, the heterogeneity of entities causes a lot difficulties for the interpretable evidence fusion reasoning process, thus a novel heterogeneous evidential chains based fusion reasoning (Hefur) method is proposed for multi-attribute group decision making. Based on the theory of evidential reasoning, the concept of evidential chain association is introduced to obtain the nearest neighbor set of distinct evidences from the data matrix of multiple decision tables. Similarity matrices are derived from data tables, and positive semi-definite matrix quadratic optimization model is built to share, sharing the experience knowledge of the group decision-making experts. Using the Dempster's quadrature rule, the rationality of the belief integrating is verified in the interpretable reasoning process with heterogeneous entities, and the combined belief is obtained from nearest neighbor evidences for each data table using the evidence fusion rules. Moreover, the validity is verified for dealing with the harmonic information inconsistence of the multi-heterogeneous data sources. Numerical experiments on the heart disease diagnosis with entity heterogeneity illustrate the feasibility and rationality of the proposed method.

Key words Entity heterogeneity, evidential chain association, similarity matrix, fusion reasoning, wisdom of crowds

Citation Shen Jiang, Yu Hai-Yan, Xu Man. Heterogeneous evidence chains based fusion reasoning for multi-attribute group decision making. *Acta Automatica Sinica*, 2015, **41**(4): 832−842

数据异构性是影响多属性群决策的可解释性推理性能的关键, 广泛存在于工程实践和管理中. 例如, 同一组织机构的不同部门之间, 不同的组织机构或合作伙伴之间, 共享和交换各自收集、存储的异构数据, 特别是企业兼并重组后, 需要进行数据集成或信息融合. 又如, 在医疗决策中, 美国麻省理工学院 (Massachusetts Institute of Technology, MIT) 等基于 Web 的复杂生理信号和生物医学信号研究资源平台, 提供多参数重症监护室的临床决策数据库[1], 各个决策数据中异构性数据表分享了大量专家的经验知识. 这些数据源自不同的关系数据库、不同水平的专家经验知识、多传感器感知数据集等, 数据实体因不同的特征属性和关系而具有异构性 (又称异质性). 目前数据异构性问题的研究已经成为多属性群决策分析领域中的热点[2−3].

随着多传感器感知信息积累, 大数据的分块存储和处理, 以及新出现的案例和决策规则知识日益增长, 决策者所面临异构性数据处理工作日趋复杂, 大多数传统的异构数据推理方法假设输入的数据集从单个数据表中获得, 没有考虑数据的实体异构性问题, 而实际决策时往往需要从多个关系数据库获

收稿日期 2014-09-09 录用日期 2014-12-12
Manuscript received September 9, 2014; accepted December 12, 2014
国家自然科学基金 (71171143, 71201087, 71271122), 天津市科技支撑计划重点项目 (13ZCZDSF01900), 中央高校基本科研业务费专项资金资助项目 (NKZXB1458) 资助
Supported by National Natural Science Foundation of China (71171143, 71201087, 71271122), Key Project of Science and Technology Supporting Program in Tianjin (13ZCZDSF01900), and Fundamental Research Funds for the Central Universities (NKZXB1458)
本文责任编委 王红卫
Recommended by Associate Editor WANG Hong-Wei
1. 天津大学管理与经济学部 天津 300072 2. 南开大学工业工程系 天津 300457
1. College of Management and Economics, Tianjin University, Tianjin 300072 2. Department of Industrial Engineering, Nankai University, Tianjin 300457

取推理的相关知识, 并且一个实体在数据库中可能会因首次出现或完全匹配的结果不存在[4], 而依据单个数据源推理的类别结果未考虑到从不同数据集中推理收集的多种证据的群体智慧[5]. 实际需要根据多个相似实体之间的共享信息积累证据进行决策. 与将单个数据集作为决策信息源的推理问题相比, 对多数据表中异构实体数据推理问题更加复杂. 首先, 每个信息源提供的决策数据表可靠性、证据参考价值不同, 这些数据集中的异构性实体对查询案例的关联作用也不同, 需要在推理结果中体现各个关联信息源的可信度; 其次, 多决策数据表特别是大数据分块推理[6−7] 中, 需要构建一个异构数据源的融合推理方法, 按照一定的融合规则综合决策推论的输出, 解决各信息表对推理结果存在的不一致性, 使得其性能优于依据大多数单数据表的推理结果. 因此, 研究多源信息异构性实体决策信息中的融合推理问题具有挑战性和实际价值.

针对多属性群决策中异构性数据融合推理, 本文所关注的两个要点为: 1) 依据从关系型数据库的异构数据中提取的决策相关属性, 识别各个异构性实体与哪些证据关联程度最紧密, 其推论更为可信. 并将这一识别模式用于推导新的测试数据集, 而这些测试数据集的标识暂不可知, 或因难以获取, 或仅能在决策之后才能获得; 2) 融合多个数据集所得到的推理证据, 获取推论的可信度分布, 消除多数据表提供的证据信息对查询案例推论存在的不一致性, 以提供更加精确的方案.

针对多属性群决策中异构性数据融合推理的相关研究主要有两类. 1) 研究的是全域数据融合推理方法. 关系型数据库中相关的多数据表包含部分决策属性, 将这些数据表分别推理. 在涵盖所有实体的数据中, 寻求与查询案例同类别且相似性高的实体, 并将其作为证据信息; 而那些与查询案例不同类别的实体被推理出相似性低, 再将所有数据表提供的证据进行推论融合. 典型的方法包括回归模型推理方法及其改进方法, 文献 [8] 研究了多源异构关系数据库中构建基于决策树的规则推理方法, 通过回归模型选择信息增益最大的属性和跨数据库链接, 实现关联数据表的分类推理. 文献 [9] 针对数据库中不同数据表的属性和关联模式, 通过属性内隐知识的依赖关系, 传播类别标识, 但需要拓展数据库中不含有类别标识的数据表, 在其最末一列增加预测的类别标识, 进而对各个推理结果进行融合. 文献 [10] 提出多准则排序融合的证据组合方法, 以选择性融合的方式, 获取最终的组合结果. 文献 [11] 提出证据冲突衡量标准下的 Dempster-Shafer (D-S) 改进算法, 改善了处理证据冲突方面的性能. 2) 相近的研究是局域数据融合推理方法, 将关系数据库中包含多种属性集合的各个数据表融合, 形成全部决策属性组成的决策数据集, 接着在融合数据集中寻求与查询案例的近邻证据信息, 在近邻证据局域内使得与查询案例同类别的近邻证据相似性高, 同时使得与查询案例不同类别的近邻证据的相似性低, 进而在从各个数据集筛选出的所有近邻证据系列中, 融合近邻证据提供的推论信息. 典型的方法包括基于案例和规则的融合机制[11−13]、基于相似度的频率加权[12−13]、距离矩阵学习[14] 等方法. 文献 [15] 提出专家数据库系统中融合案例数据和关联规则的推理方法, 使用笛卡尔积构建联合模式关系, 将包含部分决策属性的多数据表合并, 得到涵盖决策相关的全部属性的融合数据表, 使用案例实体构建规则的前件和结论, 并将不相关的案例属性移除, 以联合模式关系和条件模式关系作为融合推理的策略. 文献 [16−18] 使用基于信念规则库的推理方法, 通过估计规则激活权重、信息源权重等参数, 降低了诊断状态转移过程中的不确定性. 文献 [19] 提出了一个证据推理的规则, 使用权值和可靠度对加权可信度分布进行扩展, 以使用 D-S 理论中的可信度分布对多条独立的证据进行融合推理. 文献 [20] 对异构数据源在模式级和案例级进行识别, 对模式级的关系 (规则) 和属性 (案例) 进行相似度匹配, 进而用分类的方法对实体进行匹配, 增强了对模式元素关系进行评估的迭代响应能力. 文献 [21] 从融合空间的角度使用案例和规则知识构建决策属性酉矩阵, 并基于奇异值分解法明确辨识数据源与查询案例之间的知识关联性, 实现推理结论可信度融合. 文献 [22] 提出了一种同时考虑证据自冲突和外部冲突的相似性测度, 结合 ISODATA 聚类方法, 利用新测度对证据源进行可信度更新. 文献 [23] 提出结合弱点关联性的概念, 提出了一种基于证据推理网络的实时网络入侵取证方法, 获取的证据链完整可信且具备实时推理的能力. 为挖掘大数据集的关联性[24], 进行实体相似性推理, 因将数据集融合成一个全域数据矩阵的方法具有一定的局限性, 针对传感器感知、分块存储的大规模决策数据的特点, 本文提出的方法属于局域数据融合推理方法.

虽然这些相关的局域数据融合推理方法为解决实体异构性的多属性群决策提供了一些思路, 但也存在一些进一步完善之处: 异构性实体之间的相似性评估方面, 文献 [12] 研究了将真实的实体在不同的数据库中使用了不同标识符的情形, 并提出了基于概率决策损失优化的实体匹配方法, 辨识多个数据库中的实体是否属于同一个. 文献 [14] 针对多个专家数据表, 研究了相似度评价中的综合距离矩阵并进行分类推理. 文献 [25] 将相似性加权的频率和先验概率结合得到后验概率, 对部分相似实体的推理预测. 文献 [26] 针对一个客观实体在不同的数据库中记录不同时, 使用概率分布从这些可能值集合

中选择最好的值, 并指出这些概率能对给定的决策问题最小化错误推理损失. 证据融合的参数确定方面, 文献 [14] 使用基于专家知识的距离测度学习推理实体之间的特征相似性, 提出综合距离集成方法, 将从每个数据矩阵中获得可区分的近邻信息及单个优化的距离矩阵, 并构建基于加权参数融合各个距离矩阵的优化问题, 求解全域一致性的权重矩阵, 其特点是共享多个数据矩阵的推理结论而不共享隐性的证据数据. 类似的数据源权重处理比较经典的方法是基于民主投票的方式, 通过大多数的决策规则推理预测出决策类别标识, 其使用的条件是各个信息源 (如决策者提供的案例) 权重是一致的. 此外, 相关的方法还包括使用互信息的特征选择方法[27], 估计信息源的属性权重, 将各关联数据表融合后可以消除冗余性, 提升推理效率. 可见, 在多属性群决策的局域数据融合推理中, 针对一个数据表的异构性实体在多个其他数据表中并行匹配研究方面还不深, 本文给出一个实体异构性下证据链融合推理的多属性群决策方法. 从多个决策数据表的异构数据中获得可区分的近邻证据集合, 通过相似度矩阵进行优化推理, 并使用证据融合规则实现来自各个数据表的推理结论融合. 通过可解释性的融合推理方法, 提升异构实体下多属性群决策的信息共享能力及决策鲁棒性.

1 问题描述

1.1 大规模数据集知识表示

多源异构信息融合过程, L 个决策方如专家个体或群体、分布式环境下大数据集等, 各提供一个信息源, 如案例数据库和决策数据表等. 信息源的序数用变量 l 表示, $1 \le l \le L$. 所有数据所形成的数据集合用 m 维数据空间 D^m 表示, $D^m \in \cup_{m\ge 1} E^m$, 其中, m 为数据空间的维度, $m \ge 1$. 第 l 个信息源的特征量用矩阵 $D_l \in E^m$ 表示, 它包含 n_l 个实体. 任务有关的决策对象的物理特征使用集合 C 表示, 在分类决策中作为类别变量, $C = \{C_r | r = 1, \cdots, K\}$, 其中, r 为决策类别的序数, K 为类别状态总数. 每个信息源由一系列证据链构成. 令 (C, R) 为命题空间, 其中可信度域 R 是一个建立在决策事件可能集合 C 上的布尔代数. 推理中的实体信息来源于所提供的数据集, 并使用证据链知识表示. 证据系列 EC 作为决策者在某时刻提供的证据链集合. 第 l 个信息源中的第 i 条证据链 $\mathcal{R}_i^l$ 表示为

$$\begin{aligned} \mathcal{R}_i^l: \quad & \text{If } \{X_{ij}^l \text{ is } x_{ij}^l | j = 1, 2, \cdots, m\} \\ & \text{Then } \{(C_{ir}^l \text{ is } c_{ir}^l, \beta_{ir}^l) | r = 1, \cdots, K\} \end{aligned} \tag{1}$$

其中, x_{ij}^l 为第 j 个前件属性 X_{ij}^l 的取值, $x_{ij}^l \in D_l$; c_{ir}^l 是其第 r 个类别属性 C_{ir}^l 的取值, $c_{ir}^l \in C$; β_{ir}^l 为推论 c_{ir}^l 的可信度, $\beta_{ir}^l \in R$.

这些多源异构信息中的证据链主要包括两类: 1) 决策者经验案例的历史决策数据积累, 以及根据环境变化进行必要修正的数据; 2) 决策者所拥有的领域知识, 如决策规则或构造的虚拟案例, 以及关系数据库的属性和关联模式等. 在数据融合领域的跟踪问题中, 证据链为可能的航迹; 在医疗诊断问题中, 证据链为医生推理决策的证据网链结构; 在基于案例和规则的融合推理中, 证据链为案例序列和规则集合.

对于查询案例集合, 所包含的特征信息用矩阵 X 表示, $X \in E^m$. 若 X 中包含 Q 个实体, 每个实体信息使用向量 $\boldsymbol{x}$ 表示, 则 $X = [\boldsymbol{x}_q]_{q=1}^{Q} = [x_{qj}]_{q=1,j=1}^{Q,m}$. 其中, q 为实体信息的序数. 常见的查询案例包括: 传递到融合中心的多源传感器感知数据、在线查询问题的数据属性值、大数据分块处理中需要推理的数据块等. 以远程诊断为例, 查询案例为状态监测、体征检查等中感知数据的特征量.

针对查询案例, 用涵盖 L 个信息源的多数据表进行推导, 寻找最近邻的证据链集合, 然后融合这些证据链获得推理结论. 多源决策异构数据中, 实体异构性作为一种特殊情形, 推理检索到的实体可能是不一致的. 实际决策工作, 如医疗诊断中同一实体为患者信息, 而其诊断状态数据或诊断信息数量有限[28], 所以更需要根据异构实体之间的共享信息进行诊断决策. 又因医师诊断水平具有异构性, 意味着不同医师对同一诊断工作具有不同的诊断水平, 等同于从不同数据库中搜到索的多个案例或在同一数据库中搜索出的多个相关案例, 它们所构成的证据序列具有不同的可信度. 实际中的决策不是将关联尺度最大的那个单一实体的结论信息直接赋予查询案例, 决策者更加倾向于将关联证据所选择出来的实体信息进行融合推理, 进而得出查询案例的结论分布特征.

1.2 可信度函数及 D-S 信息融合方法

定义 1. 设 Θ 为一有限集, Θ 中的元素是互斥的, $\Lambda \subseteq \Theta$. 在 Θ 的幂集上定义一基本信度分配函数 $m(\cdot): 2^U \to [0, 1]$ 满足: $m(\emptyset) = 0$, $\sum_{\Lambda \subseteq U} m(\Lambda) = 1$, 其中, $\emptyset$ 表示空集.

对于 $\Lambda \subseteq \Theta$, 有 $m(\Lambda) > 0$, 则 Λ 成为 m 的焦点元素或核元素, 而称 $Core = \cup_{m(\Lambda)>0} \Lambda$ 为 m 的核. 基本信度分配函数是专家给出的一种评价, 是凭经验给出的一种主观判断, $m(\Lambda)$ 表示在当前证据下对假设成立的一种信任程度.

定义 2. 对于 $\Lambda \subseteq \Theta$, 在 Θ 的幂集上, 有可信度函数 $\beta(\Lambda) = \sum \{m(B) | B \subseteq \Lambda, B \ne \emptyset\}$, 简记为

$\beta(\Lambda)$.

可信度函数是一个从可信度域映射到一个封闭实数区间的函数, 它关于包含关系单调, 下极限在 $\emptyset$ 上可达. 在可信度域的元素上, 决策者关于证据链的可信度可以根据可信度函数进行量化.

定理 1[29]**.** 假设 m_1 和 m_2 为在同一识别框架 C 下不同信息源的两个基本信度分配函数, 根据 Dempster 正交规则可得:

1) $m(\emptyset)=0$;

2) $m(A)=\frac{1}{1-\Gamma}\sum_{B\cap C=A} m_1(B)m_2(C)$.

其中, Γ 表示证据源中冲突相关的基本概率分布, $\Gamma=\sum_{B\cap C=\emptyset} m_1(B)m_2(C)>0$.

2　实体异构性下的多源证据链融合推理

为充分利用多数据源中的知识, 发挥决策中群体智慧的价值, 从各个数据集中, 通过证据链关联获取对查询案例数据集 X 最为紧密的证据系列. 使用这些数据提供的共享信息, 利用其关于查询案例的推论可信度, 通过证据融合规则实现对证据系列的融合推理, 获取查询案例的推论及其可信度分布.

2.1　异构性实体相似关联

在单个数据表中的相似推理基础上, 引入证据链关联的概念, 将多个数据表之间的实体数据关联起来, 如 X 与 D_1 的关联 (简记为 $X-D_1$)、X 与 D_2 的关联 (简记为 $X-D_2$) 等, 并在各个数据表中寻求查询案例的相似证据系列.

定义 3. 给定数据集 X、数据集 D_l 和整数 k, 查询案例 $\boldsymbol{x}_q\in X$. 将在 D_l 中获取关于 $\boldsymbol{x}_q$ 的 k 最近邻证据系列的推理过程称为证据链关联, 记为 $kNN(\boldsymbol{x}_q, D_l)$.

给定 $\Re_i^l\in D_l$, $\Re_l^S\in kNN(\boldsymbol{x}_q, D_l)$, 并且 $\Re_{i'}^l\in D_l-kNN(\boldsymbol{x}_q, D_l)$, 证据链关联的相似性测度满足:

$$\mathrm{s}_{qi'}^l(\boldsymbol{x}_q,\Re_{i'}^l)\le \mathrm{s}_{qi}^l(\boldsymbol{x}_q,\Re_i^l) \tag{2}$$

其中, $\mathrm{s}_{qi'}^l(\cdot)$ 和 $\mathrm{s}_{qi}^l(\cdot)$ 为相似性测度.

将 X 中的所有元素 $\boldsymbol{x}_q$ 与 D_l 中的 k 个最近邻实体进行关联推理, 则 X 与 D_l 的证据链关联记为 $X\propto_{kNN} D_l$, 简记为 $X\propto D_l$. 形式为

$$X\propto D_l=\left\{(\boldsymbol{x}_q,\Re_i^l)\;\middle|\;\begin{array}{l}\forall \boldsymbol{x}_q\in X,\\ \forall \Re_i^l\in kNN(\boldsymbol{x}_q,D_l)\end{array}\right\}$$

根据定义 3, kNN 关联算子是非对称的, 如 $X\propto D_l\ne D_l\propto X$, 且 $X\propto D_l$ 是 $X\times D_l$ 的一个子集. 给定 $k\le|D_l|$, $|X\propto D_l|$ 的基数是 $k\times|X|$.

关于定义 3 中的相似性测度, 在证据链关联中常使用关联尺度, 实现查询案例 $\boldsymbol{x}_q$ 和各个数据源 D_l 中实体之间的多维属性变量关联. 关联尺度是一个量化测量知识组合紧密性的矩阵. 常用的关联尺度包括相关系数、距离尺度、关联系数或者概率相似度.

将关联矩阵记为 Ξ, $\Xi=[s_{qi}^l]_{Q\times n_l}$, 其中的元素 $s_{qi}^l(\boldsymbol{x}_q,\Re_i^l)$: $\mathbf{R}^m\times\mathbf{R}^m\to\mathbf{R}_{++}$, s_{qi}^l 是查询案例 $\boldsymbol{x}_q$ 与数据源 D_l 中的第 i 实体的相似性度量. 这个相似性尺度是异构数据组合 $(\boldsymbol{x}_q,\Re_i^l)$ 相近程度在数量上的度量. 在知识库中关联度量有多种方法, 针对 $X\propto D_l$, 这里使用指数型相似度:

$$s_{qi}^l(\cdot)=\exp\left(-\sum_{j=1}^m w_j(\boldsymbol{x}_i^l-\boldsymbol{x}_q)^2\right)=\exp\left(-(\boldsymbol{x}_i^l-\boldsymbol{x}_q)^{\mathrm{T}}W(\boldsymbol{x}_i^l-\boldsymbol{x}_q)\right) \tag{3}$$

其中, $\boldsymbol{x}_q$ 和 $\boldsymbol{x}_i^l$ 分别表示 $\boldsymbol{x}_q$ 和 $\Re_i^l$ 的观测值向量, W 为对称半正定矩阵, $W\in E^{m\times m}$.

对于融合推理决策信息的使用, 还需要将证据链中符号型标识的定性推论与其数值型的可信度分布建立逻辑关系. 这里引入可信度序关系, 用以使用具有一致性的可信度函数进行多属性群决策的融合推理. 将多源信息获取的各个局域证据进行融合, 获得一个全域的推论. 针对各个信息源, 利用其中与查询案例关联最紧密的信息进行共享.

定义 4[30]**.** 命题空间 (C, R) 中存在可信度函数 β, $\forall c_{r_1}, c_{r_2}\in C$, 其对应的可信度分别为 β_{r_1} 和 β_{r_2}. 可信度序关系 $\succ$ 满足:

$$c_{r_1}\succ c_{r_2}\leftrightarrow\beta_{r_1}>\beta_{r_2} \tag{4}$$

因此, 在多源异构性实体信息的决策环境中可将定性的类别辨识问题转化定量的可信度推理. 在 D_l 中任意证据链 $\mathcal{R}_i^l$ 推论的 β_{ir}^l $(r=1,2)$ 与另一证据链 $\mathcal{R}_{i'r}^l$ 的类别标识的 $\beta_{i'r}^l$ 的可信度序关系一致, 则满足: 当 $\beta_{i,r=1}^l>\beta_{i,r=2}^l$, 则 $\beta_{i',r=1}^l>\beta_{i',r=2}^l$; 当 $\beta_{i,r=1}^l\le\beta_{i,r=2}^l$, 则 $\beta_{i',r=1}^l\le\beta_{i',r=2}^l$.

定义 5. 在 $X\propto D_l$ 中, 将 D_l 中与 $\boldsymbol{x}_q$ 具有一致可信度序关系的 $|N_o^l(i)|-$最近邻集称为同构近邻, 记为 $N_o^l(i)$. 在 $X\propto D_l$ 中, 将 D_l 中与 $\boldsymbol{x}_q$ 不具有一致可信度序关系的 $|N_e^l(i)|-$最近邻集称为异构近邻, 记为 $N_e^l(i)$.

可见, 同构近邻是具有一致的类别标识的证据系列; 异构近邻是具有不一致的类别标识的证据系列.

对于查询案例, 给定 0−1 决策变量 δ_{qi}^l, 对于 s_{qi}^l, 当 $\boldsymbol{x}_i^l\in\{N_o^l(i),N_e^l(i)\}$, $\delta_{qi}^l=1$; 否则, $\delta_{qi}^l=0$. 使用同构近邻和异构近邻所形成的两个子集的信息矩阵, 构建实体异构性多源数据集的证据链融合推

理模型 (Hefur):

$$\beta_q^l(r)=\frac{\sum\limits_{i=1,2,\cdots,n_l} s_{qi}^l\cdot\delta_{qi}^l\cdot\beta_{ir}^l}{\sum\limits_{i=1,2,\cdots,n_l} s_{qi}^l},\quad r=1,2 \tag{5}$$

其中, s_{qi}^l 为相似性测度, β_{ir}^l 为对应的证据链的先验可信度. s_{qi}^l 中的参数将在第 2.2 节中进行优化学习.

2.2 证据链融合推理参数优化学习

在同构近邻 $N_o^l(i)$ 和异构近邻 $N_e^l(i)$ 的子集中, 使用式 (3), 推导出的相似度分别为

$$s_{qi}^l(o)=\exp(-(\boldsymbol{x}_i^l-\boldsymbol{x}_{i'}^q)^{\mathrm{T}}W(\boldsymbol{x}_i^l-\boldsymbol{x}_{i'}^q)) \tag{6}$$

$$s_{qi}^l(e)=\exp(-(\boldsymbol{x}_i^l-\boldsymbol{x}_{i'}^q)^{\mathrm{T}}W(\boldsymbol{x}_i^l-\boldsymbol{x}_{i'}^q)) \tag{7}$$

其中, $\boldsymbol{x}_i^l\in N_o^l(i)$, $\boldsymbol{x}_i^l\in N_e^l(i)$.

推理辨识框架为

$$J^l=\sum_{q=1}^{Q}(\ln s_{qi}^l(e)-\ln s_{qi}^l(o)) \tag{8}$$

这使得同构实体的数据关系紧密而异构实体的数据关系疏远.

因为 $W\in E^{m\times m}$ 为对称半正定矩阵, 因此采用不完全 Cholesky 分解因式分解:

$$W=ww^{\mathrm{T}} \tag{9}$$

其中, w 为一个下三角矩阵, w^{T} 为 w 的转置矩阵.

则 J^l 可以转化为

$$J^l=\mathrm{tr}(w^{\mathrm{T}}(S_{qi}^l(e)-S_{qi}^l(o))w) \tag{10}$$

其中, $\mathrm{tr}(\cdot)$ 为矩阵的迹; $S_{qi}^l(e)$ 为异构测度矩阵, $S_{qi}^l(e)=\sum_{q=1}^{Q}(-(\boldsymbol{x}_i^l-\boldsymbol{x}_{i'}^q)(\boldsymbol{x}_i^l-\boldsymbol{x}_{i'}^q)^{\mathrm{T}})$, $\boldsymbol{x}_i^l\in N_o^l(i)$; $S_{qi}^l(o)$ 为同构测度矩阵, $S_{qi}^l(o)=\sum_{q=1}^{Q}(-(\boldsymbol{x}_i^l-\boldsymbol{x}_{i'}^q)(\boldsymbol{x}_i^l-\boldsymbol{x}_{i'}^q)^{\mathrm{T}})$, $\boldsymbol{x}_i^l\in N_o^l(i)$.

因此, 对 Hefur 模型中的参数进行学习优化:

$$\begin{aligned}&\max_{w} J^l=\mathrm{tr}(w^{\mathrm{T}}(S_{qi}^l(e)-S_{qi}^l(o))w)\\&\text{s.t. } w^{\mathrm{T}}w=I\end{aligned} \tag{11}$$

其中, $(S_{qi}^l(e)-S_{qi}^l(o))$ 为 $S_{qi}^l(e)$ 与 $S_{qi}^l(o)$ 所构成的判别矩阵. 目标函数反映了能使得决策分类标识能力最大化, 这一推理模型尽可能使得查询案例的最近邻同类实例关联紧密, 异构实体疏远. 正交性约束 $w^{\mathrm{T}}w=I$ 意味着 w 为数据源信息相关联的方差阵, 对信息源矩阵中的特征信息进行选择和加权, 消除冗余性信息. 在查询案例的结论推理过程中, $X\neq D_l$, 但当训练学习参数时, $X=D_l$, 形成监督学习. 因此, 这一基于指数型相似度的参数学习问题转化为二次优化问题.

在各个数据集中, 参数优化学习过程意味着提炼各专家经验的隐性知识. 与单个数据表的推理相区别的是, 对多数据表中异构性实体之间的相似度推理, 分别完成这些优化学习过程, 并得出参数的局域解; 而不需要一次性学习优化得出参数的全域解. 这避免了将所有的这些数据表进行整合, 因为实际决策如大数据分布式数据表、群决策的各个数据表中分块的数据映射、融合更加有效.

2.3 多数据集中证据链融合

多源异构群决策因数据集的实体异构性, 查询案例依据单个最相似的证据链得出的推理结果解释能力有限, 或因受到决策者决策水平、数据的非平衡性等因素影响, 各个数据集在证据链融合推理中会存在不一致或冲突的情形, 所以有必要针对各个数据集的推理结果进行可信度融合. 通过多个决策数据集共同提供证据来积累推论可信度, 利用各个数据表多个近邻的优选证据链提供的共享信息, 以增强推理过程的解释能力. 为将 L 个数据集 $D_1, D_2, \cdots, D_L$ $(L\geq 2)$ 的推理结果有效集成, 提出多证据链可信度融合定理, 将各个数据集中的 $kNN(\boldsymbol{x}_q, D_1)$、$kNN(\boldsymbol{x}_q, D_2)$ 等进行融合.

定理 2. 在 (C, R) 上, 对于二元分类决策, $C=\{C_r|r=1,2\}$. $\beta_i^l(C_r^i)$ 是第 l 个决策方数据表提供的近邻证据系列的局域融合可信度, $\beta_i^l(C_r^i)\in\{\beta_{i,1}^l,\beta_{i,2}^l\}$. 对于查询案例 $\boldsymbol{x}_q$, 在多决策数据集中的全域可信度融合规则为

$$\beta^q(C_r^q)=\frac{1}{1-\Gamma}\sum_{\cap_{l=1}^{L}C_r^i=C_r^q}\left(\prod_{l=1}^{L}\beta_i^l(C_r^i)\right) \tag{12}$$

其中, $\Gamma=\sum_{\cap_{l=1}^{L}C_r^i=\emptyset}(\prod_{l=1}^{L}\beta_i^l(C_r^i))$.

证明. $\beta_i^l(C_r^i)$ 是第 l 个决策方数据表提供的近邻证据系列的局域融合可信度, 由式 (4) 知, $\beta_q^l(r)=(\sum_i \mathrm{s}_{qi}^l\cdot\delta_{qi}^l\cdot\beta_{ir}^l)/\sum_i \mathrm{s}_{qi}^l, r=1,2$. 当 $\boldsymbol{x}_i^l\in\{N_o^l(i),N_e^l(i)\}$ 时, δ_{qi}^l=1. 因为二元分类决策 $C=\{C_k|k=1,2\}$, 则幂集 $2^C=\{\emptyset,\{C_1\},\{C_2\},\{C_1,C_2\}\}$. 根据证据融合理论, 基本可信度分布使用映射函数 $m(\cdot)\rightarrow[0,1]$ 表示, 并且满足的性质包括: $m(A)\geq 0, A\in 2^{\boldsymbol{C}}; m(\emptyset)=0$; $\sum_{A\in 2^C}m(A)=1$. 又因可信度函数 $\beta(A)=\sum_{A\in 2^C, C_i\subset A}m(C_i)$. 对于二元分类决策, 如果融合决策信息完备, 则 $m(\{C_1,C_2\})$=0; $\beta(\{C_k\})=m(\{C_k\})$, 且 $\sum_{k=1,2}\beta(\{C_k\})=\sum_{A\in 2^{\boldsymbol{C}}}m(A)=1$. 因此, 对于 $\beta(\{C_k\})$, 适用于 Dempster 融合公式的条件, 使用定理 1, 可推导出式 (12). □

特别地, 给定所有数据库的集合 $D^*\in\cup_{m\geq 1}E^m$, 第 l 个决策方数据特征矩阵 $D_l\in E^m$, 对于 L 个决策方的信息源, 当 $L=2$ 时, 证据链可

信度融合规则为

$$\beta^q(C_r^q)=\frac{1}{1-\Gamma}\sum_{\cap_{r=1}^2 C_r^i=C_r^q}(\beta_i^1(C_r^i)\cdot\beta_i^2(C_r^i)) \quad (13)$$

其中, $\Gamma=\sum_{\cap_{l=1}^2 C_r^i=\emptyset}(\beta_i^1(C_r^i)\cdot\beta_i^2(C_r^i))$.

因此, 本方法对案例特征的融合推理过程使用了其他信息源中紧邻证据系列的共享信息, 不再提供单一的点估计, 而是以候选证据链信度为依据, 为群决策问题提供决策序列. 决策方推理的结论是概率分布集合, 增强了推理结论的可解释能力.

3　模型稳定性分析与求解步骤

3.1　稳定性分析

定理 3. 给定目标函数 $J^l=\mathrm{tr}(w^{\mathrm{T}}(S_{qi}^l(e)-S_{qi}^l(o))w)$ 的最优解为 $\boldsymbol{W}^*=[w_1^*,\cdots,w_m^*]$, 条件为 $w^{\mathrm{T}}w=I$, 其中, $S_{qi}^l(e)\in E^{m\times m}$ 和 $S_{qi}^l(o)\in E^{m\times m}$ 分别为从决策方 l 中获取的关于查询案例 $\boldsymbol{x}_q$ 的同构相似度矩阵和异构相似度矩阵. 给定判别矩阵 $(S_{qi}^l(e)-S_{qi}^l(o))\in E^{m\times m}$, 其特征值 $\sigma_1^l>\sigma_2^l>\cdots>\sigma_m^l$, 则 $\boldsymbol{W}^*=[w_1^*,\cdots,w_m^*]$ 为对应的满足正交变换的特征向量, 且 $\max\mathrm{tr}(w^{\mathrm{T}}(S_{qi}^l(e)-S_{qi}^l(o))w)=\sum_{i=1}^k\sigma_i^l$ $(k\le m)$, 其中, σ_i^l 为判别矩阵的特征值.

证明. 定义第 l 决策方的同构邻接矩阵 $w_l^o\in E^{n_l\times n_l}$ 和异构邻接矩阵 $w_l^e\in E^{n_l\times n_l}$ 的 (i,j) 元素分别为

$$w_l^o(i,j)=\begin{cases}1,\ \boldsymbol{x}_i^q\ \mathrm{adj}\ \boldsymbol{x}_j^l,\ \boldsymbol{x}_i^l\in_o^l(i)\\ 0,\ \boldsymbol{x}_i^q\ \mathrm{nadj}\ \boldsymbol{x}_j^l,\ \boldsymbol{x}_i^l\notin_e^l(i)\end{cases}$$

$$w_l^e(i,j)=\begin{cases}1,\ \boldsymbol{x}_i^q\ \mathrm{adj}\ \boldsymbol{x}_j^l,\ \boldsymbol{x}_i^l\in_e^l(i)\\ 0,\ \boldsymbol{x}_i^q\ \mathrm{nadj}\ \boldsymbol{x}_j^l,\ \boldsymbol{x}_i^l\notin_e^l(i)\end{cases}$$

其中, n_l 为第 l 决策方的数据集的实体数量. 在各个训练数据库中分别计算 $\boldsymbol{W}$.

设 $G_l^o=\mathrm{diag}\{\sum_j w_l^o(1,j),\cdots,\sum_j w_l^o(n,j)\}$ 是 $n\times n$ 对角同构邻接矩阵, 其在对角线上的第 i 个元素等于 w_l^o 的第 i 行的总和. 则定义第 l 决策方的同构拉普拉斯算子为 $L_l^o=G_l^o-W_l^o$, $L_l^o\in E^{n_l\times n_l}$. 类似地, 定义异构拉普拉斯算子为 $L_l^e=G_l^e-W_l^e$, $L_l^e\in E^{n_l\times n_l}$. 因此目标函数式 (11) 可写成

$$\begin{aligned}&\mathrm{tr}(w^{\mathrm{T}}(S_{qi}^l(e)-S_{qi}^l(o))w)=\\ &\quad\mathrm{tr}(w^{\mathrm{T}}X(L_l^o-L_l^e)X^{\mathrm{T}}w)=\\ &\quad\sum_{i=1}^k w_i^{\mathrm{T}}X_lL_l^o-L_l^e{X_l}^{\mathrm{T}}w_i=\\ &\quad\sum_{i=1}^k w_i^{\mathrm{T}}X_lL_l{X_l}^{\mathrm{T}}w_i\end{aligned}$$

其中, L_l 是差分拉普拉斯矩阵, X_l 为第 l 决策方的实体集合. 因此,

$$\max\mathrm{tr}(w^{\mathrm{T}}(S_{qi}^l(e)-S_{qi}^l(o))w)=\sum_{i=1}^k\sigma_i \qquad \square$$

定理 3 论证了在判别矩阵稳定性条件下, 通过优化模型式 (11) 可求解得出融合推理的参数矩阵 w. w 经正交变换处理后得到 $\boldsymbol{W}^*$. 其计算复杂度在最坏的情形下达 $\mathrm{O}(m^3)$. 在推理中所选择的近邻证据链比较稀疏, 在同一条件下可采用简化的近似估算过程, 以降低计算的复杂度. 通过计算判别矩阵的主特征值 σ_1^l 相应的主特征向量, 并正规化特征向量 w_1^*, 将其分量 w_j 作为对应元素的权值. 权值中接近于零的 $m-k$ 个分量所对应的属性数据视为冗余信息, 在推理过程中不予计算, 其他的 k 个属性数据构成融合推理的优化特征集, 这些特征权值对应的分量构成向量 w_k.

因此, 相似度矩阵的特征值使得优化模型具有稳定性, 证据可信度的凸组合通过多个数据集的级联, 推导出来的可信度不会在这些数据集单独推导出的可信度的连接区间之外. 优化参数能够使得集成的推论可信度对类别辨识能力更强, 推理过程利用了决策的群体智慧, 使得融合推理模型对于未标识类别的查询案例具有决策鲁棒性.

3.2　基于 MapReduce 的推理过程

针对查询案例序列, 在各个数据集中启发式检索与当前情形最相似的证据链集. MapReduce 技术框架作为面向大数据分析和处理的并行计算模型, 采用元数据集中管理、数据块分散存储的模式[31]. 本方法基于 MapReduce 的框架进行融合推理. 利用所提出的模型, 通过查询案例序列信息与已有数据集的证据链关联, 形成融合推理步骤.

步骤 1. Map 阶段证据链映射, 输入键值对 (证据编码, 证据链信息向量). 针对查询案例中的每一个实体 $\boldsymbol{x}_q\in X$, 映射函数 $map\ (\cdot)$ 对从一个数据表 X 到另一个数据表 D_l 中的每一个实体 $\mathcal{R}_i^l\in D_l$ 赋予一个键. 进而将大规模的数据集划分为 L 个数据块, 依据这个键将 D^* 划分成不相交子集, 如 $D^*=\cup_{1\le l\le L}D_l$. 对异构信息提取特征属性, 对于不同类型的数据, 使用离散化、符号属性数值化、归一化等方法处理.

输出键值对 (数据表编码, (证据编码, 证据链信息向量)).

步骤 2. 分块信息传递过程中, 针对 X 中的实体 $\boldsymbol{x}_q$, 不使用任何的修剪规则, X 的整个集合都被发送到每个融合器中, 以与 D_l 中数据进行相似度推理.

在数据重新组合 (Shuffling) 中, 每个 D_l 被传递到一个融合器中. 因此 D_l 中的实体将被复制和

传递到多个融合器中.

步骤 3. Reduce (融合) 阶段, 输入键值对 (数据表编码, 证据链信息向量).

针对查询案例, 融合机将传递来的证据信息执行 kNN 关联. $X \propto D^* = X \propto \cup_{1\le l\le L} D_l$.

结合优化特征集及权值 w_k, 使用指数型相似度式 (3) 对查询案例与证据链属性进行关联匹配并获取 s_{qi}^l.

在每个数据库中, 精炼证据链集合, 并使用使用式 (4) 计算出 β_{ik}^l.

推理机 l ($l = 1, \cdots, L$): 使用证据链 $\mathcal{R}_i^l$ 进行关联推理, $i = 1, \cdots, n_l$, 得到:

$$temp_i^{(l)} = kNN(\boldsymbol{x}_i, D_l)$$

$$\beta_q^l(r) = \frac{\sum\limits_i \mathrm{s}_{qi}^l \cdot \delta_{qi}^l \cdot \beta_{ir}^l}{\sum\limits_i \mathrm{s}_{qi}^l}$$

输出键值对 (查询编码, $(kNN(\boldsymbol{x}_i, D_l), \beta_q^l(r))$.

步骤 4. 推论信息分享, 将 D_l 中关于 $\boldsymbol{x}_i$ 的 $|N_o^l(i)| + |N_e^l(i)|$ 个最近邻实体传递到同一个融合器, 并将它们赋予 $\boldsymbol{x}_i$ 同样的键.

融合 $kNN(\boldsymbol{x}_i, D_l)$, 使用定理 2 的式 (12), 计算融合可信度值 $\beta^q(C_r^q)$, 并更新结论信息 $(C_r^q, \beta^q(C_r^q))$.

通过信息融合推理的逆变换过程分享证据链. 数据重新组合的复杂度为 $|D^*| + L \cdot |X|$.

如果证据链所提供的方案在过去是成功的, 则直接使用这一证据链所对应的方案, 并可根据当前的状况做适当的调整; 如果没有检索到历史证据链, 则根据专家经验或领域知识规则给出一个当前的方案, 并记录该方案的决策结果, 将其记录入案. 输入的案例序列如果矩阵, 则经证据链关联后, 将优化的结论分享与每个输入案例, 并与领域知识 (或专家) 得出的病理结论比对, 检验模型性能. 新的案例或规则, 可固化知识形成新的关联证据链, 以多次利用证据数据提升决策价值.

4 应用实例分析

决策数据集为从大规模的医疗电子病历 (EHRs)、传感器感知信息以及专家对样本做类别标识的经验知识等信息源中截取的分块数据. 这些多源异构决策信息源的数据集覆盖一系列代表单个实体 (如医疗患者) 的事件. 为讨论患者的病情存在诊断和治疗的难题, 医疗专家组从不同科室调来大量的拥有丰富经验知识专家进行决策, 这些专家拥有相当于一个独立数据集的知识库. 使用本文方法进行临床决策支持, 通过具有实体异构性的相似病案信息共享, 进行融合推理决策, 并在具有不同数量的同构实体和异构实体的决策数据表上, 对相关方法及推理结果进行性能比较.

4.1 实验平台和决策数据集

实验的操作系统是 Ubuntu10.04, 数据库管理系统是 PostgreSQL8.4.8, 处理器配置为 Intel Core2 P8400 (2.93 GHz, 2 G). 考虑到大规模数据的分布式数据库推理融合问题, 实验数据使用 UC Irvine 决策数据库[32] 中的 Heart (Cleveland) 数据集 D_1 和 Heart (Hungarian) 数据集 D_2 作为群决策训练数据集, Heart (Long Beach VA) 作为测试集. 使用的数据集信息表, 如表 1 所示.

表 1 实验使用的 UCI 数据集信息表

Table 1 Experimental information of the UCI data sets

多源数据	数据	属性数	类别标识	类分布
D_1	Heart (Cleveland)	13	Present; Absent	164/139
D_2	Heart (Hungarian)	13	Present; Absent	188/106
X	Heart (Long Beach VA)	13	?	51/149

这些决策数据集来自于不同的信息源 (关联数据库), 它们通过匹配方式与特定患者关联并完成时间对准. 将包含患者的识别信息 (姓名、医疗病历编号) 的波形数据文件、生理数据记录 (以案例 ID 为索引) 与相应的临床信息记录匹配.

在 Heart (Cleveland) 数据集中, 303 个连续的病人案例所记录的实体均龄 54 岁, 68 %为男性, 心脏病的患者比例为 54.13 %. 在 Heart (Hungarian) 数据集中, 294 个连续的病案所记录的实体中, 心脏病的患者比例为 63.95 %. 数据集信息包括所有患者病历和生理检查、静息心电图和化验记录等多源异构数据. 这些数据集的获取所使用的多源传感器包括静脉压检测仪、血清蛋白测量仪、血糖测量仪、心率测试仪和心电监护仪等, 其心脏病数据记录有多个特征属性, 包括患者的心电图、脉搏波、血压、呼吸波、液晶屏上起搏操作同步记录、药物种类、给药剂量等 76 种特征. 不同属性特征在心脏病急救决策中发挥的作用不同, 其中一些属性特征对知识推理具有重要作用, 本文使用常用于诊断推理的 13 个特征. 样本空间中每个样本有一个由专家根据经验或医疗领域知识给出的类别标识, 即这些数据集被分离为四种类型的心脏病和没有心脏病, 按二元分类将 CHD 划分为 Present 和 Absent, 分别记为 C_1 和 C_2.

这些关系数据融合过程经过 2 个阶段: 1) 将来自检测仪 (传感器) 生成的数据记录中的姓名和医疗记录编号 (可获得的准确记录过的) 与系统中的临床数据记录的对应部分相匹配; 2) 包括从测试数据集,

如在线监测的检测数据中的生理趋势信息与临床信息系统中的监护人员检验过的生命体征信息相匹配, 寻找最近邻的证据支持. 经过数据库融合过程, 实现了患者的多源异构数据集中管理, 供异构性数据的进一步融合推理.

4.2　预处理与诊断推理

数据预处理过程中, 对于训练集中的 6 个缺失数据被丢弃, 27 个争议数据被修改. 对逻辑布尔型属性和描述型属性进行符号化处理, 将所有属性的各种取值映射为符号, 对于描述型属性, 根据取值区间分别映射, 如将属性 Cp 的取值 typical angina 、atypical angina、non-anginal pain 和 asymptomatic 分别映射为 1、2、3 和 4. 使用这一数据集中的 190 个和 100 个样本分别作为训练集和测试集.

使用前文中参数学习优化方法及定理 3, 通过计算判别矩阵的主特征向量并正规化, 将其分量作为对应元素的权值. 对 Heart (Cleveland) 数据集, 获取的优化特征集为$\{Age, Sex, Cp, BP, restECG, Thalach, Exang, Slope, Thal\}$, 这 9 个特征对应的 w_k 中的分量分别为 [0.0743, 0.0105, 0.2342, 0.0111, 0.0352, 0.1030, 0.1577, 0.1437, 0.2303]. 对 Heart (Hungarian) 数据集, 获取的优化特征集中这 9 个特征对应的 w_k 中的分量分别为 [0.1735, 0.0588, 0.0588, 0.1471,0.0588, 0.2353, 0.0912, 0.1176, 0.0589].

在数据集 D_1 中, 以其中一个案例数据为例. 实体信息如下: Age 年龄 ($Year$) 为 57, 性别为男, 胸痛类型 (Cp) 为 2 (atypical angina), 血压 (Systolic Blood Pressure, BP) 为 124 mmHg; 安静时的心电图结果 ($Restecg$) 为 0 (normal); 最高心率 ($Thalach$) 为 141; 是否运动导致心绞痛 ($Exang$) 为 0 (no); 峰值 ST 倾斜角度 ($Slope$) 为 1 (向上倾斜) 和心跳情况 ($Thal$) 为 7 (可逆缺陷). 将这一多源异构信息源获取数据转化为证据链, 为

$$
\begin{aligned}
R_1^1\text{: If } & Age \text{ is } 57 \wedge Sex \text{ is } 1 \wedge Cp \text{ is } 2 \wedge \\
& BP \text{ is } 124 \wedge restECG \text{ is } 0 \wedge \\
& Thalach \text{ is } 141 \wedge Exang \text{ is } 0 \wedge \\
& Slope \text{ is } 1 \wedge Thal \text{ is } 7 \\
\text{Then } & (CHD \text{ is } Present, \beta_1^i = 100\,\%), \\
& (CHD \text{ is } Absent, \beta_2^i = 0\,\%)
\end{aligned}
$$

证据链所表示的传感器感知的信息或电子病历的体征变量, 常按照心脏病诊断临床路径获取.

类似地, 对于数据集 D_2, 将其多源异构信息源获取的一个证据链实例为

$$
\begin{aligned}
R_1^2\text{: If } & Age \text{ is } 41 \wedge Sex \text{ is } 2 \wedge Cp \text{ is } 1 \wedge \\
& BP \text{ is } 128 \wedge restECG \text{ is } 2 \wedge \\
& Thalach \text{ is } 137 \wedge Exang \text{ is } 1 \wedge \\
& Slope \text{ is } 2 \wedge Thal \text{ is } 4 \\
\text{Then } & (CHD \text{ is } Present, \beta_1^i = 0\,\%), \\
& (CHD \text{ is } Absent, \beta_2^i = 100\,\%)
\end{aligned}
$$

使用 Key 和 Value 表示多源数据表的关联 Map Out 表和使用 D-S 规则推导的测试案例的融合结果 Reduce Output 表输出的字段和取值, 如表 2 和表 3 所示.

表 2 中, 查询案例在数据表 D_1 中得出 $|N_o^l(i)|+|N_e^l(i)|=3$ 时的近邻证据链为 EC 列的 EC_1、EC_4 和 EC_{52}, 所对应的 δ_{qi}^l 都取值为 1, 识别各个异构性实体相关联的最可靠的证据集合. 进而使用式 (4) 计算出近邻证据链对查询案例的相似度分别 81.47 %、85.07 % 和 68.37 %. 依据数据表训练数据的类别标记, 当样本取值为 Present 时, 将 β_1^i 和 β_2^i 分别赋值为 100 % 和 0; 当样本取值为 Absent 时, 将 β_1^i 和 β_2^i 分别赋值为 0 和 100 %.

类似地, 可得出这一查询案例在数据表 D_2 中的近邻证据链、相似度和各个可信度.

表 2　多源数据表的关联 Map Out 表

Table 2　Map Out table associated with multi-datasets

Key	Value					
$LineID$	$Dataset$	EC	δ_{qi}^l	s (%)	β_1^i (%)	β_2^i (%)
1	D_1	EC_1	$\delta_{1,1}^1=1$	81.47	100	0
2	D_1	EC_4	$\delta_{1,4}=1$	85.07	0	100
3	D_1	EC_{52}	$\delta_{1,5}=1$	68.37	0	100
4	D_2	EC_2	$\delta_{2,1}=1$	90.83	100	0
5	D_2	EC_{41}	$\delta_{2,4}=1$	82.56	100	0
6	D_2	EC_{67}	$\delta_{2,6}=1$	78.57	0	100
⋮	⋮	⋮	⋮	⋮	⋮	⋮

表 3　测试案例的 Reduce Output 结果表

Table 3　Reduce Output result table of the testing cases

Key	Value		
D	Nearest ECs	β_1^i (%)	β_2^i (%)
D_1	$\delta_{1,1}^1=1,\ \delta_{1,4}=1, \delta_{1,52}=1$	71.20	28.80
D_2	$\delta_{1,2}^1=1, \delta_{2,41}=1;\ \ \delta_{2,67}=1;$	68.81	31.19
⋮	⋮	⋮	⋮

表 3 中, 依据从关系型数据表获取的近邻异构证据链集合, 通过式 (4) 计算出各个数据表的集成

可信度, 如查询案例从数据集 D_1 中获得的集成可信度分别为 71.2 % 和 28.8 %, 并将这些可信度分布用于进一步计算多数据表的融合信度. 针对查询案例在 D_1 和 D_2 中分别得出的可信度, 使用定理 2 的融合规则, 得出 $D_1 \vee D_2$ 的融合信度为: $\beta_1^i = 84.51\,\%$; $\beta_2^i = 15.49\,\%$.

4.3 实验结果

使用 Hefur 模型求解, 计算测试数据集中的查询案例 1 ($X_1 \in X$) 在不同的同构实体和异构实体下的融合信度, 如图 1 所示. 针对查询案例集合 X, 使用本方法在单个数据集下的推理准确度 D_1 (Hefur) 与 D_2 (Hefur), 以及本方法 Hefur 与 Comdi 方法[14] 在多个测试数据集 ($D = D_1 \vee D_2$) 下的推理准确度, 结果比较如图 2 所示.

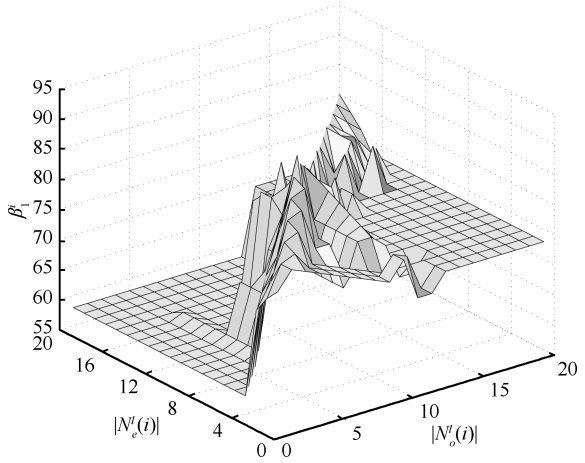

图 1 对测试数据集中查询案例 X_1 推理出的融合可信度

Fig. 1 Combing belief of the case X_1 from the testing data with the reasoning method

图 1 中 x 和 y 轴分别表示样本数据的同构近邻证据和异构近邻证据的数量 $|N_o^l(i)|$ 和 $|N_e^l(i)|$; z 轴表示数据的融合可信度 β_1^i. 在训练集 D_1 和 D_2 的优化过程中 $|N_o^l(i)| = |N_o^l(i)|$ 时, 查询案例的推论可信度取最大值. 且随着 $|N_o^l(i)|$ 的增加, 查询案例的可信度在一定范围内增加, 而随着 $|N_e^l(i)|$ 的增加, 查询案例的可信度在一定范围内降低. 在针对查询案例, 根据其信度和类别逆向推理, 可在证据链关联矩阵中查询中最相关的证据链集合, 并将对应的信息分享给诊断决策用户.

图 2 中横轴表示使用的测试数据集及对应的方法, 即 D_1 (Hefur)、D_2 (Hefur)、D (Hefur) 和 D (Comdi), 纵轴表示方法的准确度. 准确度[21] 为 $Acc = (TP + TA)/(TP + TA + FP + FA)$, 其中, TP、FN 表示为查询案例的实际类别 C_1 分别被推理为 C_1 和 C_2 的样本数; FP、TN 分别表示查询案例实际类别为 C_2 而分别被推理为 C_1 和 C_2 的样本数.

图 2 中的结果表明了查询案例从训练集中分别获取 3 个最近邻证据链时, 在不同数据集或不同方法下得到的推理准确度. 从比较结果可以看出, D (Hefur) 的准确度均值为 89.05 %, 比 D_1 (Hefur) 的准确度均值 84.27 % 和 D_2 (Hefur) 的 82.69 % 更高, 并且其方差也更小, 即 D (Hefur) 的总体性能更好. 这是因为所提出的方法在数据集上实现了更大规模的决策信息共享. D_1 (Hefur) 的准确度均值比 D_2 (Hefur) 的性能更好, 是因为后者的数据非均衡性 (188/106) 比前者的数据非均衡性 (164/139) 更高. 同时, D (Hefur) 在准确度均值比 Comdi 方法在同一规模的数据集上的准确度均值 87.84 % 更高, 并且方差也更小, 这是因为所提出的方法通过多数据集的融合规则获得了推理的近邻证据集及其可信度分布, 消除了多数据表提供的证据信息对查询案例推论可能存在的不一致性而提供了更加准确的方案.

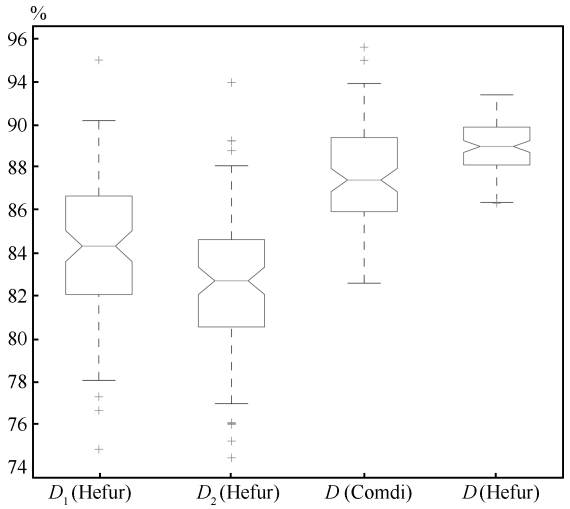

图 2 所提出方法在多决策表下的推理准确度比较

Fig. 2 Accuracy comparison of the reasoning methods on multiple decision datasets

从决策信息结构来看, 文献 [14] 所提出 Comdi 方法使用多个决策信息源的全域数据, 通过相似度评价获得综合距离矩阵, 以距离最近邻的局域数据作为推理证据对查询案例进行判别分类. 与此相区别的是, 本文所提出的 Hefur 方法针对多数据源中各决策方提供的局域数据, 使用相似度矩阵对分块数据的决策参数进行优选学习, 实现查询案例在多数据表中的并行匹配和证据融合, 并以可信度序关系作为定性分析到定量判别分类依据, 因此提升了可解释性推理的准确性和鲁棒性.

5 结论

异构数据融合特别是在大数据或分布式存储等新兴的群决策数据处理中, 日益成为工程实践和管

理中多属性群决策的焦点问题. 由于单个决策者或单个数据库的知识有限性, 需要对多数据表信息进行异构数据融合. 然而, 现有的多个关系数据库的融合主要集中在数据表的同质性合并及融合推理上, 对群决策下多数据表中的异构性实体数据的相似性推理研究不深, 因此本文提出了实体异构性下的证据链融合推理多属性群决策方法. 与采用单数据表的信息源融合推理方法相比, 本方法针对查询案例在各个数据表中相似匹配的异构性实体数据, 分享多源决策的近邻证据链, 进而各数据表提供的可信度信息, 而不需要构建大规模数据的稀疏矩阵 (如 Comdi 方法中的综合距离矩阵[14], 基于频率相似度加权的概率方法中的联合数据矩阵[13]). 另外, 该方法最大限度的考虑到了多数据表之间的异构性: 1) 各个数据表中的实体异构性通过求解基于相似度矩阵的二次优化求解特征值, 获得了最佳的属性权重, 使得与查询案例的同构近邻和异构近邻快速获得. 2) 针对多数据表之间可能存在的证据可信度不一致性, 使用证据融合规则将各个数据表的结论进行融合, 其可解释性的融合推理过程提升了异构实体数据之间的信息共享能力. 对于异构数据的多源信息融合推理, 时空异构环境下信息融合推理动态过程, 包括证据链前件的推导、部分信息下的动态推理等, 作为进一步的研究方向. 另外, 对应于多数据表中各证据链类别的可信度的精确获取, 也可作为后续研究的内容.

References

1 Scott D, Lee J, Silva I, Park S, Moody G, Celi L, Mark R G. Accessing the public MIMIC-II intensive care relational database for clinical research. *BMC Medical Informatics and Decision Making*, 2013, **13**: 9

2 Scott M, Boardman R P, Reed P A, Cox S J. Managing heterogeneous datasets. *Information Systems*, 2014, **44**: 34−53

3 Hoffmann S, Fischbeck P, Krupnick A, McWilliams M. Elicitation from Large, Heterogeneous expert panels: using multiple uncertainty measures to characterize information quality for decision analysis. *Decision Analysis*, 2007, **4**(2): 91−109

4 Krishnan R, Li X P, Steier D, Zhao L. On heterogeneous database retrieval: a cognitively guided approach. *Information Systems Research*, 2001, **12**(3): 286−301

5 Baron J, Mellers B A, Tetlock P E, Stone E, Ungar L. Two reasons to make aggregated probability forecasts more extreme. *Decision Analysis*, 2014, **11**(2): 133−145

6 O′Leary D E. Artificial intelligence and big data. *IEEE Intelligent Systems*, 2013, **28**(2): 96−99

7 Fan J Q, Han F, Liu H. Challenges of big data analysis. *National Science Review*, 2014, **12**(1): 293−314

8 Mehenni T, Moussaoui A. Data mining from multiple heterogeneous relational databases using decision tree classification. *Pattern Recognition Letters*, 2012, **33**(13): 1768−1775

9 Manjunath G, Narasimha Murty M, Sitaram D. Combining heterogeneous classifiers for relational databases. *Pattern Recognition*, 2013, **46**(1): 317−324

10 Yang Yi, Han De-Qiang, Han Chong-Zhao. Evidence combination based on multi-criteria rank-level fusion. *Acta Automatica Sinica*, 2012, **38**(5): 823−831
(杨艺, 韩德强, 韩崇昭. 基于多准则排序融合的证据组合方法. 自动化学报, 2012, **38**(5): 823−831)

11 Hu Chang-Hua, Si Xiao-Sheng, Zhou Zhi-Jie, Wang Peng. An improved D-S algorithm under the new measure criteria of evidence conflict. *Acta Electronica Sinica*, 2009, **37**(7): 1578−1583
(胡昌华, 司小胜, 周志杰, 王鹏. 新的证据冲突衡量标准下的 D-S 改进算法. 电子学报, 2009, **37**(7): 1578−1583)

12 Dey D, Sarkar S, De P. A probabilistic decision model for entity matching in heterogeneous databases. *Management Science*, 1998, **44**(10): 1379−1395

13 Billot A, Gilboa I, Schmeidler D, Samet D. Probabilities as similarity-weighted frequencies. *Econometrica*, 2005, **73**(4): 1125−1136

14 Wang F, Sun J, Ebadollahi S. Composite distance metric integration by leveraging multiple experts′ inputs and its application in patient similarity assessment. *Statistical Analysis and Data Mining*, 2012, **5**(1): 54−69

15 Segev A, Zhao J L. Rule management in expert database systems. *Management Science*, 1994, **40**(6): 685−707

16 Yang J B, Liu J, Wang J, Sii H, Wang H. 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

17 Wang J Q, Zhang H Y. Multicriteria decision-making approach based on atanassov′s intuitionistic fuzzy sets with incomplete certain information on weights. *IEEE Transactions on Fuzzy Systems*, 2013, **21**(3): 510−515

18 Wang J Q, Nie R R, Zhang H Y, Chen X H. Intuitionistic fuzzy multi-criteria decision-making method based on evidential reasoning. *Applied Soft Computing*, 2013, **13**(4): 1823−1831

19 Yang J B, Xu D L. Evidential reasoning rule for evidence combination. *Artificial Intelligence*, 2013, **205**: 1−29

20 Zhao H, Ram S. Combining schema and instance information for integrating heterogeneous data sources. *Data & Knowledge Engineering*, 2007, **61**(2): 281−303

21 Xu M, Yu H Y, Shen J. New algorithm for CBR-RBR fusion with robust thresholds. *Chinese Journal of Mechanical Engineering*, 2012, **25**(6): 1255−1263

22 Li Xin-De, Wang Feng-Yu. A method of evidence reasoning based on the ISODATA clustering and improved similarity Measure, *Acta Automatica Sinica*, 2015, **41**(3): 575−590
(李新德, 王丰羽. 一种基于 ISODATA 聚类和改进相似度的证据推理方法, 自动化学报, 2015, **41**(3): 575−590)

23 Tian Zhi-Hong, Yu Xiang-Zhan, Zhang Hong-Li, Fang Bin-Xing. A real-time network intrusion forensics method based on evidence reasoning network. *Chinese Journal of Computers*, 2014, **37**(5): 1184−1194
(田志宏, 余翔湛, 张宏莉, 方滨兴. 基于证据推理网络的实时网络入侵取证方法. 计算机学报, 2014, **37**(5): 1184−1194)

24 Reshef D N, Reshef Y A, Finucane H K, Grossman S R, McVean G, Turnbaugh P J, Lander E S, Mitzenmacher M, Sabeti P C. Detecting novel associations in large data sets. *Science*, 2011, **334**(6062): 1518−1524

25 Bordley R F. Using Bayes′ rule to update an event′s probabilities based on the outcomes of partially similar events. *Decision Analysis*, 2011, **8**(2): 117−127

26 Jiang Z R, Sarkar S, De P, Dey D. A framework for reconciling attribute values from multiple data sources. *Management Science*, 2007, **53**(12): 1946−1963

27 Xu M, Yu H-Y, Shen J. New approach to eliminate structural redundancy in case resource pools using alpha mutual information. *Journal of Systems Engineering and Electronics*, 2013, **24**(4): 625−633

28 Yang Jin-Feng, Yu Qiu-Bin, Guan Yi, Jiang Zhi-Peng. An overview of research on electronic medical record oriented named entity recognition and entity relation extraction. *Acta Automatica Sinica*, 2014, **40**(8): 1537−1562
(杨锦锋, 于秋滨, 关毅, 蒋志鹏. 电子病历命名实体识别和实体关系抽取研究综述. 自动化学报, 2014, **40**(8): 1537−1562)

29 Basir O, Yuan X H. Engine fault diagnosis based on multi-sensor information fusion using Dempster-Shafer evidence theory. *Information Fusion*, 2007, **8**(4): 379−386

30 Wong S K M, Lingras P. Representation of qualitative user preference by quantitative belief functions. *IEEE Transactions on Knowledge and Data Engineering*, 1994, **6**(1): 72−78

31 Xue Yong-Jian, Ni Zhi-Wei. Research of large scale manifold learning based on MapReduce. *Systems Engineering Theory & Practice*, 2014, **34**(S1): 151−157
(薛永坚, 倪志伟. 基于 MapReduce 的大规模数据集流形学习降维研究. 系统工程理论实践, 2014, **34**(S1): 151−157)

32 Asuncion A, Newman D. UCI machine learning repository. [Online], available: http://www.ics.uci.edu/mlearn/ML Repository.html, October 28, 2010

沈 江 天津大学管理与经济学部教授. 主要研究方向为信息融合, 多传感器数据获取和群决策.
E-mail: motoshen@163.com
(**SHEN Jiang** Professor at the College of Management and Economics, Tianjin University. His research interest covers information fusion, multi-sensor data acquisition, and group decision-making.)

余海燕 天津大学管理与经济学部博士研究生. 2009 年获得南京邮电大学经济与管理学院学士学位. 主要研究方向为证据推理, 医疗数据挖掘和基于相似推理. E-mail: yhy188@tju.edu.cn
(**YU Hai-Yan** Ph. D. candidate at the College of Management and Economics, Tianjin University. He received his bachelor degree from Nanjing University of Posts and Telecommunications in 2009. His research interest covers evidential reasoning, medical data mining, and similarity-based reasoning.)

徐 曼 南开大学工业工程系讲师, 2011 年获得天津大学博士学位. 主要研究方向为基于规则推理, 信息融合和医疗诊断决策. 本文通信作者.
E-mail: twinklexu@163.com
(**XU Man** Lecturer in the Department of Industrial Engineering, Nankai University. She received her Ph. D. degree from Tianjin University in 2011. Her research interest covers rule-based reasoning, information fusion, and medical diagnosis decision. Corresponding author of this paper.)