分 类 号 TP18 密级 公开

收藏编号\_\_\_\_\_\_\_\_\_\_\_ 学号\_N180327111\_\_

学校代码 10386 编号\_\_\_\_\_\_\_\_\_\_\_\_\_

工程硕士标题

(应用研究)

**基于析取范式的置信规则库优化方法及**

**应用研究**

|  |  |
| --- | --- |
| 工程领域： | 软件工程 |
| 研究方向： | 智能决策与专家系统 |
| 研究生姓名： | 刘永裕 |
| 指导教师、职称： | 傅仰耿 副教授 |
| 协助导师、职称： |  |
| 所在学院： | 数学与计算机科学学院 |
| 答辩委员会主席： |  |

二〇二一 年 一 月

**一 遵守学术行为规范承诺**

本人已熟知并愿意自觉遵守《福州大学研究生和导师学术行为规范实施办法》和《福州大学关于加强研究生毕业与学位论文质量管理的规定》的所有内容，承诺所提交的毕业和学位论文是终稿，不存在学术造假或学术不端行为，且论文的纸质版与电子版内容完全一致。

**二 独创性声明**

本人声明所提交的论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得福州大学或其他教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。本人完全意识到本声明的法律结果由本人承担。

**三 关于论文使用授权的说明**

本人完全了解福州大学有关保留使用学位论文的规定，即：学校有权保留送交论文的复印件，允许论文被查阅和借阅；学校可以公布论文的全部或部分内容，可以采用影印、缩印或其他复制手段保存论文。（保密的论文在解密后应遵守此规定）

**本学位论文属于（必须在以下相应方框内打“√”，否则一律按“非保密论文”处理）：**

1、保密论文： □本学位论文属于保密，在 年解密后适用本授权书。

2、非保密论文：□本学位论文不属于保密范围，适用本授权书。

研究生本人签名： 签字日期：20 年 月 日

研究生导师签名： 签字日期：20 年 月 日

基于析取范式的置信规则库优化方法及应用研究

摘要

Yang等人在2006年首次提出了基于证据推理的置信规则库推理方法，将置信规则结构分为合取范式和析取范式。析取范式的连接方式能够有效地降低系统的规则数量，减小系统参数规模，从本质上解决置信规则库的“组合爆炸”问题。到目前为止对于基于析取范式的置信规则库的已有许多相关研究，但仍存在众多问题，尤其对于数据缺失问题仍缺少足够的研究。首先，基于析取范式的置信规则库推理方法的准确率依赖于训练数据集的质量，当数据集不完整时将导致训练出的模型不够准确；其次，基于析取范式的置信规则库专家系统是一个能够处理复杂问题的决策系统，但对于不完整输入信息的决策问题，仍缺乏相关研究，除了采用传统的填补数据方法外，缺少较好的处理方法。针对上述问题，本文展开了课题研究工作，具体如下：

(1) 针对现有的基于析取范式的置信规则库在不完整数据集上存在的难以训练参数和推理不准确的问题，提出基于扩展规则的析取范式置信规则库优化方法。该方法首先利用历史数据生成规则，并针对训练数据集中的不完整数据，构建不完整的规则表示，然后引入衰减因子对不完整的规则权重进行调节，使该规则能够更合理地参与信息融合过程。最后，在五个公共测试数据集上进行对比实验，证明了本文方法的有效性。说明该方法能够在不增加系统复杂度的条件下，完成对完整数据集或不完整数据集的训练，并保证一定的推理准确率。

(2) 针对现有的基于析取范式的置信规则库对于不完整信息决策问题，提出基于模糊关联规则的析取范式置信规则库优化方法。首先该方法对完整的训练数据通过隶属函数建立模糊置信表示库，然后对输入的数据进行模糊化表示，并通过模糊置信表示库对未缺失的数据信息构建关联规则，推断出缺失数据的置信分布情况，最后将得到的完整数据输入训练好的析取范式置信规则库系统中进行推理。通过在数据集上的对比实验，验证本方法能够通过已有的信息去获取缺失的数据信息，使基于析取范式的置信规则库对不完整输入信息能有较好的推理准确率。

(3) 利用本课题提出的基于析取范式的置信规则库优化方法建立教师教学水平评估模型。测试数据集来自福州大学2018学年01学期的真实数据。本应用首先使用ER算法获取教师教学评分情况，然后结合基于析取范式的置信规则库优化方法建立教师教学水平评估模型，最后该模型与支持向量回归模型、决策回归树和多层感知器进行对比，从而证明本文方法在完整数据集和不完整数据集上具有较好的评估性能。

关键词：置信规则库；析取范式；数据缺失；不完整信息；

The optimization method and application of disjunctive belief rule base

Abstract

Yang et al. firstly proposed the belief rule-based inference methodology using the evidential reasoning approach in 2006 and divided the structure of belief rule into conjunctive form and disjunctive form. The disjunctive belief rule-based system can effectively reduce the number of rules in rule base and the size of system parameters, which solves the “combination explosion” problem of belief rule-based system in essence. However, up to now, there are many problems that occur in the research on the disjunctive belief rule-based system, especially the lacking of sufficient research on data missing. First of all, the reasoning accuracy of disjunctive belief rule-based system depends on the quality of training dataset. When the dataset is incomplete, the trained model will be inaccurate. Secondly, the disjunctive belief rule-based system is capable of handling complex problems, but there is still a lack of relevant research on the decision problems with incomplete input information. Besides the conventional data filling method, a better processing method is required. In view of the above problems, this paper conducts research work specifically listed as follows:

(1) Focusing on the difficulty of parameter training and inaccuracy of reasoning on the existing disjunctive belief rule-based system applied on incomplete datasets, an optimization structure of disjunctive belief rule-based system using extended rules is proposed. Firstly, this method uses historical data to generate rules, constructs incomplete rules according to the incomplete data. Secondly, this paper introduces an attenuation factor to adjust the weight of incomplete rules, so that the rules can participate in the information aggregation process more reasonably. Finally, the UCI datasets are selected to make comparison experiments and demonstrate the effectiveness of this method. It shows that this method can process the training of complete or incomplete datasets without increasing the complexity of system, and ensure the reasoning accuracy.

(2) Based on the existing disjunctive belief rule-based system for decision-making problems with incomplete information, an optimization method based on fuzzy association rules is proposed. Firstly, the fuzzy belief representation base is established for the complete training data through the membership function. Then the fuzzy representation of the input data is carried out, association rules are constructed for the dismissing data information through the fuzzy belief representation base, and the belief distribution of the missing data is inferred. Finally, the complete data is input into the disjunctive belief rule-based system for reasoning. In test datasets, validation of this method can effectively by the existing information to get the missing data information, disjunctive belief rule-based system can obtain a good reasoning accuracy on incomplete input information.

(3) Using the optimization of disjunctive belief rule-based system proposed in this paper, a teacher teaching level evaluation model is established. The test data set used is derived from the real data of Fuzhou University in the first semester of 2018. The system uses the ER algorithm for teachers’ teaching evaluation at first and then combined with the proposed disjunctive belief rule-based system optimization method to establish the teaching level evaluation model. Finally, the model is compared with the support vector regression model, decision regression tree, and multilayer perceptron model, respectively. It is proved that the proposed method has a good performance in evaluating both complete and incomplete datasets.

Key words: Belief Rule-Base, Disjunctive Normal Form, Missing Data, Incomplete Information

目录

[摘要 I](#_Toc61273094)

[Abstract II](#_Toc61273095)

[第一章 引言 1](#_Toc61273096)

[1.1 课题背景与意义 1](#_Toc61273097)

[1.2 国内外研究现况 2](#_Toc61273098)

[1.3 论文主要内容 4](#_Toc61273099)

[1.4 论文组织结构 5](#_Toc61273100)

[第二章 基于析取范式的置信规则库相关知识 7](#_Toc61273101)

[2.1 引言 7](#_Toc61273102)

[2.2 基于合取范式的置信规则库 7](#_Toc61273103)

[2.2.1 基于合取范式的规则结构 7](#_Toc61273104)

[2.2.2 规则推理方法 8](#_Toc61273105)

[2.3 基于析取范式的置信规则库 11](#_Toc61273106)

[2.3.1 DBRB规则结构 11](#_Toc61273107)

[2.3.2 规则库构建方式 11](#_Toc61273108)

[2.3.3 DBRB规则推理方法 12](#_Toc61273109)

[2.4 本章小结 13](#_Toc61273110)

[第三章 基于扩展规则的DBRB优化方法 14](#_Toc61273111)

[3.1 引言 14](#_Toc61273112)

[3.2 EBRB专家系统 14](#_Toc61273113)

[3.2.1 规则表示 15](#_Toc61273114)

[3.2.2 构建规则库 15](#_Toc61273115)

[3.2.3 规则推理 16](#_Toc61273116)

[3.3 基于扩展规则的DBRB 17](#_Toc61273117)

[3.3.1 BRB中的不完整数据集问题 17](#_Toc61273118)

[3.3.2 规则表示 18](#_Toc61273119)

[3.3.3 激活权重计算公式 18](#_Toc61273120)

[3.3.4 前提属性参考值个数的可变 18](#_Toc61273121)

[3.3.5 对缺失数据的处理 20](#_Toc61273122)

[3.4 算法复杂度分析 22](#_Toc61273123)

[3.5 实验分析 22](#_Toc61273124)

[3.5.1 与其他EBRB的对比实验 22](#_Toc61273125)

[3.5.2 不完整数据集的对比实验 25](#_Toc61273126)

[3.6 本章小结 30](#_Toc61273127)

[第四章 基于模糊关联规则的DBRB优化方法 31](#_Toc61273128)

[4.1 引言 31](#_Toc61273129)

[4.2 构建模糊置信表示库及DBRB规则库 31](#_Toc61273130)

[4.2.1 前提属性参考值的模糊表示 32](#_Toc61273131)

[4.2.2 构建模糊置信表示库 33](#_Toc61273132)

[4.2.3 基于差分进化的DBRB 34](#_Toc61273133)

[4.3 关联规则推断缺失值方法 36](#_Toc61273134)

[4.3.1 关联规则 36](#_Toc61273135)

[4.3.2 缺失数据的推断方法 37](#_Toc61273136)

[4.4 算法复杂度分析 38](#_Toc61273137)

[4.5 实验分析 39](#_Toc61273138)

[4.5.1 方法验证实验 39](#_Toc61273139)

[4.5.2 与DBRB方法进行比较 41](#_Toc61273140)

[4.5.3 与其他方法进行对比 43](#_Toc61273141)

[4.6 本章小结 44](#_Toc61273142)

[第五章 基于DBRB的教师教学水平评估应用 45](#_Toc61273143)

[5.1 引言 45](#_Toc61273144)

[5.2 基于DBRB的教师教学水平评估模型 45](#_Toc61273145)

[5.2.1 采用ER算法融合教师测评信息 45](#_Toc61273146)

[5.2.2 DBRB优化方法应用 46](#_Toc61273147)

[5.3 实例测试 47](#_Toc61273148)

[5.3.1 教师测评得分对比实验 47](#_Toc61273149)

[5.3.2 基于DBRB的评估方法与其他方法的比较 48](#_Toc61273150)

[5.4 本章小结 52](#_Toc61273151)

[总结与展望 53](#_Toc61273152)

[参考文献 55](#_Toc61273153)

[致谢 59](#_Toc61273154)

[个人简历 60](#_Toc61273155)

[在学期间的研究成果及发表的学术论文 61](#_Toc61273156)

# 第一章 引言

## 课题背景与意义

专家系统诞生于1965年，在几十年间飞速地发展并取得了极大的进步。目前专家系统在理论和技术都已趋近于成熟，并在人工智能领域、商业领域和工业领域发挥了重要的作用[1]。专家系统是通过模拟专家解决专业领域问题所使用的逻辑或推理方法的一种决策系统，通过使用规则、启发算法和其他技术象征性方式去表示知识，它能够根据问题的领域或解决方法创建知识库，然后通过推理引擎利用知识库对问题进行有效的整合、判断或选择和处理[2]。规则表示的专家系统是以人类的历史经验和储备知识作为基础，通过IF-THEN规则结构进行表示，并构建相应的知识库。该结构的专家系统更符合人类的思考方式，同时Chomsky等[3]和Sum等[4]研究了规则表示的知识形式，认为其它知识表示形式均可转化为基于规则的表示方式，因此规则库专家系统已在决策支持系统和人工智能领域取得了广泛的应用。

专家系统结构主要包含有知识库、推理机、解释器、数据库和人机接口等，其中知识库和推理机是专家系统的主要组件。知识库结构由知识表示方式所决定，一般基于规则库的专家系统采用IF-THEN规则形式进行知识表示。同时知识库不仅存储领域知识、常识性知识，还需要存储专家经验总结知识。推理机是知识库的处理器，根据用户观察到的数据利用知识库中的知识按照一定的推理策略去求解用户提出的问题。通过知识库和推理机的相互组合，使得专家系统能够从专家知识角度对用户问题进行解答，但由于人类知识存在含糊性和不精确性或不完整性所引起的不确定性[5][6]，会使得数据含有模糊不准确性和概率不确定性。

许多学者已经针对不确定性问题提出各种解决方法，最常用的方法包括有基于概率论的Bayes方法、Dempster-Shafer证据理论、模糊集理论和神经网络等。但每个方法都有各自的局限性，例如Bayes概率论方法侧重于处理定量的数据，无法处理定性数据，且该方法需要事先确定数据的分布，而数据的分布形式往往是无法确定的[7]；D-S证据理论可以处理概率不确定性，同时它能够处理定量和定性的数据，但无法处理模糊不确定性；模糊理论能很好的处理模糊不确定性，但无法较好的处理概率不确定性[8]。

在2006年，Yang等[9]提出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)。该方法采用IF-THEN规则表示专家知识，并引入了置信框架对传统的IF-THEN规则结构进行扩充得到置信规则库(Belief Rule Base，BRB)。对于知识推理部分，引入了模糊理论、D-S证据推理方法和决策理论[10][11][12]来处理模糊不确定性和概率不确定性问题，使得BRB具有更好的非线性数据表征能力。以RIMER为基础的专家系统统称为置信规则库系统[1]，它本质上是由一系列的置信规则所构成，并通过证据推理(Evidential Reasoning，ER)算法[12]来实现知识推理。它的推理流程大致分为四步：首先是用户输入数据，然后将数据与规则库中的规则进行匹配并收集，再使用ER算法对收集到的规则进行信息合成，最后向用户输出决策结果。BRB相比于其他的智能系统具有较少的参数设置，且过程简单具有可解释性，领域专家可以通过对置信规则的参数调整优化系统推理性能。因此，置信规则库已成功应用到各个领域，例如输油管道检漏[12]、出租车乘车概率预测[13]、工程系统安全评估[14]、武器装备军事能力评估[15]、消费者偏好预测[16]、临床诊断[17]、双边匹配问题[18]等。

Yang等[9]对BRB规则结构进行了详细的讨论并将其分为两类：合取范式、析取范式。合取范式置信规则库(Conjunctive BRB，CBRB)采用逻辑与进行规则前提属性的连接，析取范式置信规则库(Disjunctive BRB，DBRB)采用逻辑或进行前提属性的连接，除了这两种规则结构外，还可能会出现复合形式的规则结构，如

|  |  |  |
| --- | --- | --- |
|  |  | 公式(1-1) |

但这种规则最终可以分解成两个合取范式的规则结构，如

|  |  |  |
| --- | --- | --- |
|  |  | 公式(1-2) |
|  |  |

相比于DBRB，CBRB规则结构更加符合观测到的数据，且规则更容易构造，因此许多研究者着重于研究合取范式规则而忽略了析取范式规则结构。Chang等[19]在2016年针对分类问题再次提出了析取范式置信规则库，并讨论分析了DBRB与CBRB的优缺点[20]，充分挖掘了DBRB的性质，解决了BRB在多属性多参考值决策上，规则库的规模会随着前提属性个数和前提属性参考值个数的增加而呈指数形式增加，即规则数“组合爆炸”问题。DBRB已经成功应用于桥梁风险分析[21]、军事安全等级评估[22]、铁路交通安全评估[23]等。

## 国内外研究现况

早期Yang等[9]提出的BRB系统依赖于领域专家知识或决策者的历史经验，需要人为的进行参数初始化和设置。但随着应用系统越来越复杂，专家或决策者无法为系统设置最优参数。因此，Yang等[24]提出参数优化模型用来训练BRB系统参数，并使用MATLAB工具箱对参数优化模型的目标函数进行求解。Chen等[25]将属性权重加入参数训练，提高BRB系统的推理性能。然而MATLAB工具箱对过多的参数训练效果并不理想，因此Chang等[26]提出基于优化步长和梯度下降的参数学习方法，但该方法无法覆盖所有的参数且公式复杂。为了解决该问题，Wang等[27]提出了在专家干预下的差分进化算法，Su等[28]提出了使用变速粒子群优化算法训练BRB系统参数。

通过引入参数训练，使BRB系统不再完全依赖于领域专家知识经验，并提高规则库的推理精度，但参数训练需要大量繁复的迭代过程才能得到最终的推理结果，使得BRB系统的推理效率下降。一个完备的BRB系统需要对任意的输入都能有对应的输出结果，即对任意的输入都能激活规则库中的某条规则，因此BRB系统需要拥有一个完备的规则库[21]。而BRB规则库的规模与前提属性个数及其参考值个数呈指数相关，使得BRB的规则数容易出现“组合爆炸”情况，所以基于合取范式的BRB结构的研究引起了许多学者的兴趣。Chang等[29]利用多维尺度变化分析、主成分分析和灰靶理论等方法对前提属性进行属性约简，减少BRB前件结构负担。Yang等[30]使用关联系数标准差的置信规则库约简方法评估属性重要程度，依据关联权重减少规则库规模。Wang等[31]使用粗糙集理论在不依赖于规则库以外知识的情况下对属性进行约简。Zhou等[32]提出“统计效用”方法来计算具有两个先决属性的规则的效用，并以此来确定是否保留该条规则。Ken等[33]利用每个模糊分区对数据集进行划分并减少未使用的规则数。Wang等[34]引入误差分析和基于密度的空间聚类算法(DBSCAN)对规则结构进行调整，确保最后的规则库既不会出现欠拟合，也不会导致过拟合。Liu等[35]提出了扩展置信规则库方法(Extended Belief Rule Base， EBRB)，该方法在BRB的基础上，对前件引入置信分布框架，使得BRB前件能够更好地表达不确定性信息。同时EBRB能够利用历史数据生成相应的置信规则，避免了规则数“组合爆炸”的问题，开辟了BRB新的研究方向[36][37][38]。

尽管针对BRB结构优化的研究已经能在一定程度上有效地避免“组合爆炸”问题，但不可避免地会造成决策信息的损失，影响最终的推理精度。Chang等[19]采用了Yang提出的析取范式BRB规则结构，该结构能够从根本上解决BRB的“组合爆炸”问题。DBRB将规则前提属性连接方式改为“”，在规则中每个前提属性参考值只需要遍历一次，不再需要进行多次相互组合，有效地降低了规则库的数量。相比于CBRB，当规则前件相同时，DBRB只需构建更少的规则数量，能够大规模地缩减系统的冗余的规则。同时，较少的规则数代表系统具有较少的参数，在参数优化步骤只需对少量参数进行训练，比CBRB的训练时间来得更少，并且缩短了对规则的匹配时间。Chang等[19]对DBRB结构进行优化并成功地应用于分类问题中。Ye等[39]针对DBRB在分类问题上的应用，提出DBRB规则数设置准则，将分类数设置为规则数时，能使DBRB推理性能达到最佳分类能力。Chang等[40]提出了通过有限的CBRB来构造DBRB的方法，讨论了两种规则结构间的联系。Wang等[41]提出了通过不完备的BRB规则构造DBRB规则库方法，并提出能适用于多属性值的构造算法。Chang等[42]提出了使用于DBRB的联合优化方法，该方法能够有效地对非线性复杂系统进行结构调整和参数优化。

DBRB为了摆脱专家知识经验带来的局限，需要使用历史数据进行训练DBRB参数，训练数据集的质量将会影响训练的效果。可见，DBRB系统是以数据为基础并推动对问题的推理决策，而DBRB中的数据缺失问题仍还缺少充分的研究成果，因此，数据缺失问题是DBRB的研究重点之一。Chang等[22]提出使用不完整数据建立多个子合取范式BRB系统，然后通过自组织映射的方式构建DBRB系统；Wang等[41]利用同样的思想从不完整的数据中构建规则完备的DBRB系统；Yu等[50]人通过历史数据分布特征通过填补缺失数据，并采用二次ER方法获取最终的结果；Li等[51]人先使用ER方法对缺失数据进行合成，然后使用BRB系统对完整的数据集进行推理。但上述研究都存在各自的不足，因此，本将针对基于析取范式的置信规则库在数据缺失中的优化方法进行研究。

## 论文主要内容

本文主要围绕基于析取范式的置信规则库的优化方法及应用展开，首先针对析取范式置信规则库系统中的不完整数据集问题，提出基于扩展规则的析取范式置信规则库优化方法，然后针对不完整信息决策问题，利用模糊关联规则对输入的不完整数据进行处理，提出基于模糊关联规则的析取范式置信规则库优化方法，最后将基于析取范式的置信规则库模型应用于教师教学水平评估问题中。主要研究内容分以下三个方面：

第一、基于扩展规则的析取范式置信规则库推理方法。析取范式置信规则库面对不完整数据集时，含有缺失项的数据将会导致规则库参数训练出现不准确，最后的推理结果不精确。本文提出基于扩展规则的析取范式置信规则库优化方法，通过不完整数据生成不完整规则，并引入衰减因子对规则权重进行调整，减小不完整规则对推理结果的影响，使新的推理方法对不完整数据集具有良好的推理准确率。

第二、基于模糊关联规则的析取范式置信规则库方法。本节针对不完整数据决策问题，提出解决方法，首先利用高斯隶属函数替换个体匹配度计算方式，然后通过训练数据建立模糊置信表达库，通过模糊置信表达库对用户输入的不完整数据进行关联规则推断，进而使用改进的析取范式置信规则库进行推理。

第三、建立基于析取范式置信规则库的教师教学水平评估模型。该方法首先针对传统测评打分方式的不足，采用ER算法获取学生对教师测评成绩，然后建立优化的基于析取范式置的信规则库模型对教师教学水平进行评估。

## 论文组织结构

本文组织结构如图1-1所示。主要内容包含有六个章节，其中，第一第二章节主要介绍置信规则库系统背景意义以及基础理论知识；第三第四第五章节对本文工作进行详细地介绍；第六章节对工作进行总结与展望。每一章节内容主要如下：

第一章为引言部分，着重介绍置信规则库研究背景与意义，讨论置信规则库地研究方向主要分为参数学习和结构优化，介绍这些方向上国内外地研究成果，重点分析了基于析取范式置信规则库相对于合取范式置信规则库的优势与研究现况。

第二章为基础理论部分，主要介绍了置信规则库的规则表示与规则推理算法，重点描述基于析取范式的置信规则库与基于合取范式的置信规则库的区别以及基于析取范式的置信规则库的推理方式。

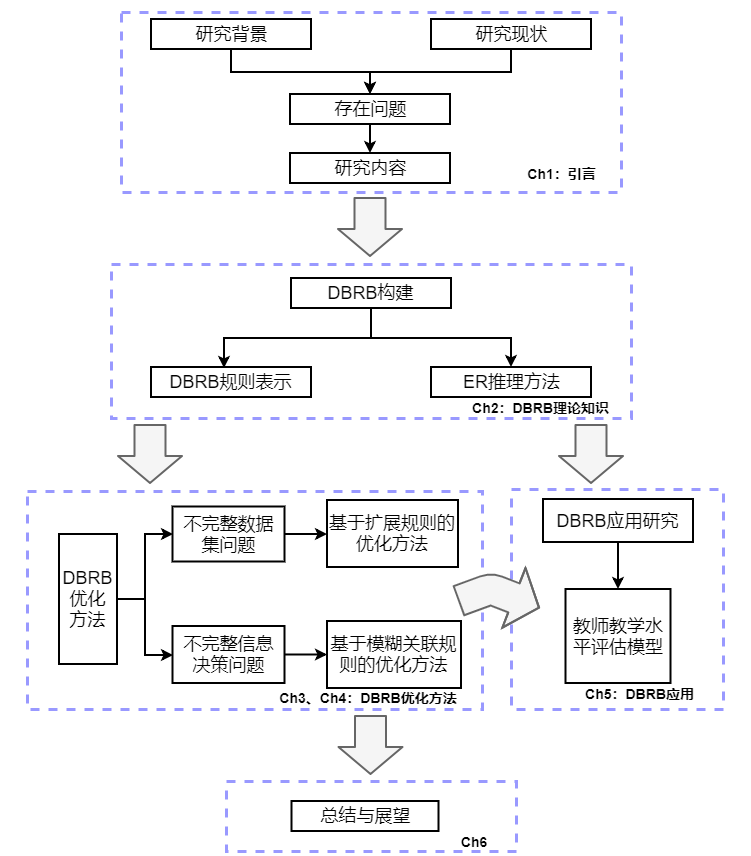


图1-1全文结构图

第三章为针对DBRB中不完整数据集的推理方法的研究，主要介绍了不完整数据集对DBRB推理的影响，通过构建基于扩展规则的析取范式置信规则库对不完整数据进行处理，并引入衰减因子对规则权重进行调节，使推理更加精确。

第四章为针对DBRB中不完整信息决策问题的研究，主要介绍了基于关联规则的DBRB对不完整信息决策问题的处理方法，首先采用高斯隶属函数计算数据的隶属度，然后通过训练数据建立模糊置信表示库，并对不完整输入数据进行模糊关联规则推断，使得DBRB能够更好地处理不完整信息决策问题。

第五章为DBRB优化方法在教师教学水平评估问题中的应用，首先使用ER算法获取教师教学水平成绩，然后采用优化的DBRB方法建立评估模型，并与支持向量回归模型、决策回归树和多层感知器进行对比，证明本文方法的有效性。

最后进行本文研究工作的总结和展望，对本文研究工作进行概括与对该领域的未来工作进行一个更深入的展望。

# 第二章 基于析取范式的置信规则库相关知识

## 引言

传统的置信规则库可以分为基于合取范式的置信规则库和基于析取范式的置信规则库，两者都是基于ER算法对规则推理并合成最后信息，但两个在规则结构上具有不同的表示方式。为了更好的介绍基于析取范式的置信规则，本章首先介绍了基于合取范式的置信规则库的规则表示和推理公式，在此基础上再介绍基于析取范式的置信规则库具有哪些不同的方面，及它的规则表示和推理流程。

## 基于合取范式的置信规则库

基于合取范式的置信规则库主要包括两个部分，分别是合取范式的置信规则结构和证据推理算法。构建一个基于规则库的决策系统需要结合领域专家知识经验或通过数据挖掘方式获取知识并使用IF-THEN规则形式进行表达，然后以此为基础根据用户的输入数据进行知识推理并反馈推理的结果。

### 基于合取范式的规则结构

基于IF-THEN的规则表示通用形式如下：

|  |  |  |
| --- | --- | --- |
|  |  | 公式(2-1) |

其中,表示第个规则的第个属性值，表示属性个数，表示规则个数；表示第条规则的推理结果，是所有可能出现的结果集。属性间的连接方式为逻辑与。

使用这种规则形式表示知识，可以通过规则前件属性匹配规则得到一个准确的结果，例如：如果自行车骑行时有噪声且刹车器比较紧那么刹车片有问题。当我们观测到上述现象时，根据规则就可以得出该去修理刹车片的结论了，这个结论也符合实际生活经验。但有时候并非如此，修车师傅可能会告知刹车片有可能坏了，但问题也可能出现在其他地方，因此，这个自行车故障预测结果是不确定的，是带有概率推测的。

可以知晓传统的IF-THEN规则无法很好的表示具有模糊、不确定、不完整的知识，鉴于此，Yang等[9]提出具有置信分布框架的规则形式，该规则结构如公式(2-2)所示。

|  |  |  |
| --- | --- | --- |
|  |  | 公式(2-2) |

其中，表示第条规则的第个属性，表示第条规则的第个属性对应的参考值，表示第个评价结果，表示第个评价结果对应的置信度，越大代表第个评价结果越可信。当时，代表第条置信规则是完整的；当时，代表第条规则是不完整的。表示第条规则的权重，表示第个属性的权重。

上述规则结构可知，带有置信分布框架的规则形式更加符合人类认知水平，同时能更好地对知识中的不确定性信息进行表示，使得该置信规则库具有更实用性和信息性的知识表示方式。但是置信规则库中的各个参数，如规则权重、属性权重和评价结果置信度的设置依赖于领域专家知识经验，因此BRB系统需要通过参数学习方法进行参数训练和更新。

### 规则推理方法

置信规则库的推理方法是基于ER算法，通过信息组合激活的规则得到带有置信分布的推理结果，根据求解问题类型输出相应的结果。该推理过程主要分为四个部分：(1)匹配规则；(2)计算激活权重；(3)ER推理；(4)输出结果。

#### 规则匹配与激活权重

BRB系统可以处理数值类型和语义类型的数据，需要将输入信息和属性参考值之间进行转换，即计算输入信息的个体匹配度。以数值类型数据为例，假设输入数据，其中表示第个属性值。利用信息转化技术，可以得到如下的分布式表示形式：

|  |  |  |
| --- | --- | --- |
|  |  | 公式(2-3) |

其中，表示第个属性的第个属性参考值，使用表示其效用值；表示输入的第个属性相对于第个属性参考值的个体匹配度，表示第个属性的参考值个数，信息转化技术公式如下：

|  |  |  |
| --- | --- | --- |
|  |  | 公式(2-4) |

那么，对于输入所激活的第条规则的激活权重计算公式如下：

|  |  |  |
| --- | --- | --- |
|  |  | 公式(2-5) |

规则激活权重，表示对各个属性权重进行归一化。需要注意到的是，公式(2-5)中对属性个体匹配度采用累乘的方式进行计算，这是因为合取式规则属性间采用"∧"表示符，各个属性间是逻辑与的关系，因此，当某个属性参考值的个体匹配度为0时，该属性参考值对应的规则便无法激活。

#### ER推理合成和输出结果

输入数据对规则库中的规则进行匹配，匹配到的规则通过2.2.2.1节公式计算规则激活权重。当激活权重大于0，该规则便被激活，即该规则对输入数据是有效的。被激活的规则将会通过ER算法进行信息合成，从而获得最后的推理结果。ER算法主要分为以下两个步骤：

首先，将激活规则的评价结果置信度转化为基本概率质量，即：

|  |  |  |
| --- | --- | --- |
|  |  | 公式(2-6) |
|  |  | 公式(2-7) |
|  |  | 公式(2-8) |
|  |  | 公式(2-9) |

其中，表示分配给评价结果的基本概率质量；表示分配给评价结果集合的基本概率质量；表示由第条规则引起的重要程度，当时，表示第条规则是绝对重要的，即；表示由第条规则评价结果的不完整性所引起的，当评价结果是完整的时，则。其中，。

然后，采用Dempster准则对激活的规则进行组合，便可得到相对评价结果集合的置信分布。公式如下：

|  |  |
| --- | --- |
|  | 公式(2-10) |
|  | 公式(2-11) |
|  | 公式(2-12) |
|  | 公式(2-13) |
|  | 公式(2-14) |
|  | 公式(2-15) |

其中，表示相对于评价结果的置信度，表示对于未分配给评价结果集合的置信度。

通过上述的迭代公式能够得到最终结果的置信分布，但上述公式较为复杂，不方便在工程中进行应用，对于此，Wang等[43]提出了解析式ER算法，该算法可以直接对BRB中所有的激活规则进行信息合成，公式如下：

|  |  |  |
| --- | --- | --- |
|  |  | 公式(2-16) |
|  |  | 公式(2-17) |

其中，为激活权重，通过公式(2-5)计算得到。上述公式是一个关于前提属性权重、规则权重和后件置信度的函数，最后该函数输出结果为

|  |  |  |
| --- | --- | --- |
|  |  | 公式(2-18) |

当求解的问题为分类问题时，选取最大置信度的评价等级结果作为输出结果：

|  |  |  |
| --- | --- | --- |
|  |  | 公式(2-19) |

当求解的问题为回归问题时，计算期望效用值作为最后输出结果：

|  |  |  |
| --- | --- | --- |
|  |  | 公式(2-20) |

## 基于析取范式的置信规则库

### DBRB规则结构

Yang等[9]在提出BRB时，讨论了规则属性连接的两种方式：“”和“”，分别是基于合取范式的置信规则库和基于析取范式的置信规则库，但是在提出的十年间，专家学者对BRB的研究都默认采用基于合取范式的规则结构，而DBRB却是无人问津，直到2016年，Chang等[19]针对分类问题提出DBRB新的激活权重计算公式，获得了显著的效果。因此DBRB开始受到学者的关注，并在工程领域进行应用。DBRB的规则表达形式如下：

|  |  |  |
| --- | --- | --- |
|  |  | 公式(2-21) |

与公式(2-2)相比较可知，DBRB除了规则前件属性连接方式不同之外，其余与CBRB具有相同的结构，其符号表示含义与CBRB是一致的，因此不再重复介绍符号含义。

### 规则库构建方式

构建基于合取范式的置信规则库需要遍历组合前件属性参考值，如图2-1所示，假设规则有个前提属性，每个前提属性有个参考值，那么最后构建的规则库含有个规则数，可见CBRB的规则数与前提属性和前提属性参考值呈指数级关系。CBRB系统面对多属性多参考值问题时，会出现规则“组合爆炸”问题。

DBRB采用””符号进行规则连接，因为规则属性间的关系从逻辑与改为逻辑或的关系，DBRB构建规则库时不需要遍历组合所有的属性参考值，而是采用线性组合规则属性的方式，此时规则库规模将由含有最多属性参考值的属性所决定，规则个数将减少为个。

假设规则库系统含10个属性，每个属性含有2个参考值，那么CBRB的规则库规模将到达条规则，而DBRB规则库仅仅需要条规则。因此可知，基于析取范式的置信规则库能够大幅度地减少规则数量，从而有效地避免了“组合爆炸”问题。

但需要注意地是，CBRB能够根据给定的属性参考值构建一个唯一的置信规则库，而对于DBRB，因为在构建规则时，只要保证每个属性参考值至少出现一次就可以，因此对于DBRB构建的规则库是不唯一的，假设规则库有两个属性，每个属性有两个参考值，分别为和，那么DBRB构建的规则有两种：和。不同的规则形式，对最后的推理结果具有一定的影响。Chang等[19]首次提出了将属性参考值作为参数，通过参数训练的方式设置规则，并取得突出的效果；Yang等[21]在Chang等[19]研究基础上，提出动态参数优化方式，使规则设置更加合理。

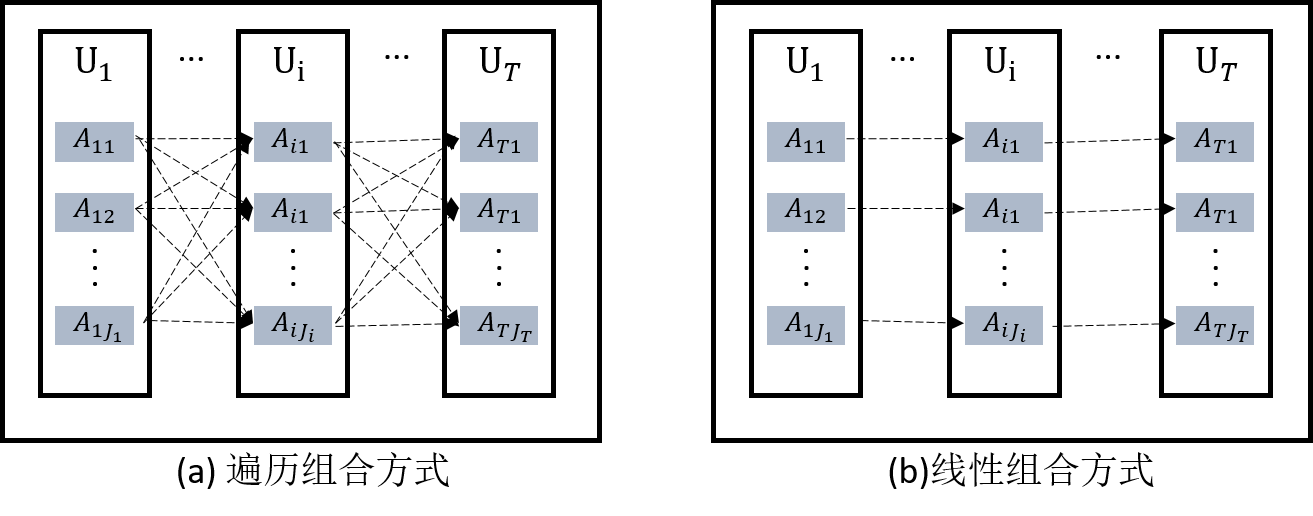


图2-1规则组合方式

### DBRB规则推理方法

DBRB规则推理方法与CBRB一致，同样采用ER算法对激活规则进行信息合成。由于DBRB规则表示与CBRB不同，因此DBRB的规则激活权重计算方式有所不同，Yang等[9]定义加权乘积和聚合函数(weighted product-sum aggregation function)计算:

|  |  |  |
| --- | --- | --- |
|  |  | 公式(2-22) |
|  |  | 公式(2-23) |
|  |  | 公式(2-24) |

其中，，为第条规则第个属性的个体匹配度。，当且时，，当时，。规则激活权重公式如下：

|  |  |  |
| --- | --- | --- |
|  |  | 公式(2-25) |

Chang等[19]提出新的激活权重计算公式：

|  |  |  |
| --- | --- | --- |
|  |  | 公式(2-26) |

该公式将公式(2-5)中属性个体匹配度的累乘换成累加，因为公式(2-5)基于概率论方法，所以使用累加形式计算激活权重更加符合基于析取范式的规则结构。

DBRB的规则推理方法中除了计算激活权重方式不同，其余与CBRB的计算方式相同，均是通过ER算法对激活的规则进行信息合成并得到最后的推理结果。

## 本章小结

本章主要介绍了基于合取范式的置信规则库和基于析取范式的置信规则库的相关知识，并重点讲解了CBRB和DBRB的规则结构、规则库构建方式和ER算法的推理过程。突出体现DBRB与CBRB的不同点与优势，为后续章节对于基于析取范式的置信规则库的研究提供理论依据。

# 第三章 基于扩展规则的DBRB优化方法

## 引言

在现实生活中，常常会由各种原因造成采集到的数据集是不完整的，而不完整数据集会对以数据为驱动的模型造成低效地分析和不准确地推理。DBRB系统通过历史数据进行参数训练，从而设置系统中的参数。但当训练数据集是不完整时，DBRB需要事先对缺失的数据进行额外的处理，这会造成系统的复杂度增加；另一方面，不完整的数据集缺失部分信息，当使用该数据集对DBRB系统进行参数训练时，容易造成最后训练得到的参数并不是最优，甚至在高缺失率的数据集中，最后的参数设置是不合理的。因此对于DBRB系统而言，如何处理不完整数据集是个重要的研究内容。

针对于DBRB中存在的不完整数据集问题，已有学者进行了一定程度的研究，Wang等[45]提出通过不完整数据集建立多个不完备CBRB系统，通过自组织映射合成完备的DBRB系统，该方法虽然降低系统的复杂性，但最后推理精度并不理想；Wang等[41]还提出通过使用等概率方法通过多个不完整CBRB构建DBRB的方法，但该方法未在实际数据集中进行验证；Chang等[22]由自组织映射思想启发，通过不完整数据集进行划分，将相似结构的数据用来构造多个不完整子规则库，并合成具有析取范式的FBRB系统，该方法能够和领域专家知识经验相互结合，成功应用于威胁等级评估，但该方法对于数据缺失模式要求较严苛，对于随机缺失的数据集，会使该系统的复杂度变大。

对BRB中的不完整数据集已有一些解决方法，但这些方法或多或少都存在自身的不足，针对这些问题提出了基于扩展规则的析取范式置信规则库方法，该方法能够处理能不同的缺失程度的数据集，也能够处理不同的缺失模式--单一属性值缺失或多个属性值缺失。为了验证所提方法的有效性，选取公共数据集中的Iris、Wine、Transfusion、Seeds和Mammographic数据集进行测试，并从实验结果可以知道本文方法不仅在完整数据集中具有良好的表现，同时在不同缺失模式和不同缺失程度下具有良好的表现。

## EBRB专家系统

EBRB由Liu等[35]提出，该BRB结构能够从已有的数据生成置信规则结构，并通过规则间的差异设置规则权重，不再需要人为设置参数和繁复的参数训练过程。本文方法借鉴于EBRB构建规则库的方式生成基于析取范式的规则结构，并采用EBRB计算公式，在此基础上提出基于扩展规则的析取范式置信规则库。

### 规则表示

EBRB的规则库同样采用IF-THEN规则结构，但在前件引入了置信分布框架，让规则属性能更好地表示不确定性信息，规则表示如下：

|  |  |  |
| --- | --- | --- |
|  |  | 公式(3-1) |

其中，表示第i个属性；表示第i个属性的第j个参考值，表示第条规则中第个属性的第个参考值的置信度；表示属性参考值的个数；其余符号和公式(2-2)一致。

### 构建规则库

EBRB的规则库规模由使用的历史数据所决定。假设有条数据，每个数据含有个属性，那么规则库中规则数量便为L。构建EBRB规则库步骤[35]如下：

步骤1：根据领域专家建议首先确定前提属性参考值个数和评价结果参考值个数，然后确定前提属性参考值和确定评价结果参考值;

步骤2：使用公式(2-4)对数据的输入和输出部分转化为置信分布的形式，如公式(2-3)；

步骤3：将L条数据都通过步骤2转化为公式(3-1)表示形式，得到初始的扩展置信规则库；

步骤4：度量规则间的不一致性程度，计算公式如下：

|  |  |  |
| --- | --- | --- |
|  |  | 公式(3-2) |
|  |  | 公式(3-3) |

得到与的一致性度量：

|  |  |  |
| --- | --- | --- |
|  |  | 公式(3-4) |

通过计算第i条规则与其他规则的不一致性，便可得到该条规则的不一致性程度：

|  |  |  |
| --- | --- | --- |
|  |  | 公式(3-5) |

通过上述步骤，可以得到各个规则的不一致性程度，通过不一致性程度设置每条规则的权重，以此来降低不一致性造成的影响：

|  |  |  |
| --- | --- | --- |
|  |  | 公式(3-6) |
|  |  | 公式(3-7) |

其中，为第条规则的权重，为常数用来调节不一致性对规则权重的影响。当规则权重没有初始数据时，则可以直接设置为：

|  |  |  |
| --- | --- | --- |
|  |  | 公式(3-8) |

### 规则推理

通过上述章节可以得到前件和后件都具有置信分布的置信规则库，接下来需要将输入数据与规则库中的规则进行匹配并激活。假设输入数据，通过公式(2-4)将该输入数据转化为置信分布的形式。计算该输入数据与第条规则中第个属性的个体匹配度：

|  |  |  |
| --- | --- | --- |
|  |  | 公式(3-9) |
|  |  | 公式(3-10) |

其中，表示输入数据的第i个属性与第k个规则的第i个属性的相似程度，表示第k条规则的第i个属性的个体匹配度。当两个数据属性的置信分布在空间中距离越近时，的取值越小，个体匹配度越大，反正，取值越大，个体匹配度越小。因为，所以需要限定最小只能取到0。

根据属性个体匹配度，计算各个规则的激活权重：

|  |  |  |
| --- | --- | --- |
|  |  | 公式(3-11) |

上述公式将公式(2-5)中的个体匹配度替换为公式(3-9)计算所得到的值，其余符号表示含义均与公式(2-5)一致。

EBRB仍采用ER算法对激活公式进行信息合成，后续流程与2.2.2.2节所述一致，最终获得的结果依据问题类型进行输出。

## 基于扩展规则的DBRB

### BRB中的不完整数据集问题

CBRB能够根据专家给定的属性参考值生成置信规则库，但对于较复杂的系统，往往需要进行参数训练，通过不断地迭代和学习，最终得到能使系统推理性能最优的参数设置。但是当数据集是不完整时，需要对数据集进行预先地处理，对缺失项进行填充，而填充算法将会影响到参数训练的效果进而影响到系统的推理精度。

Liu等[35]提出的EBRB能够不使用参数训练便能达到一个较好的推理准确性，但面对不完整的数据集，EBRB无法直接使用缺失数据生成相应的规则库。例如输入数据中第个属性是缺失项，那么在计算输入数据的置信分布时，无法对第个属性的使用公式(2-4)，此时对第个属性的置信分布是未知的，仍需要通过对缺失项进行处理。另一方面，因为EBRB规则间的属性为合取范式，不同缺失项的数据将构建不同结构的规则，将使得规则库复杂度增加。

DBRB由Yang等[9]提出，Chang在Yang的基础上改进了DBRB的激活权重并引入参数训练方法构建规则库。DBRB与CBRB面临相同的难点，不完整的规则会造成参数训练上的不准确。但对于析取范式的规则结构而言，对于不完整的数据集能更好地处理，因为对于规则属性间的权重采用逻辑与进行连接，所以一条规则被激活当且仅当该条规则中某个属性的个体匹配度不为0。对于具有缺失项的数据集，若采用不完整数据生成不完整的规则结构时，对于这些规则中其他未缺失属性的计算并没有太大的影响。例如上述的数据中，使用该数据生成规则时，尽管第t个属性具有缺失值，但””连接方式使得该规则得其他属性仍能够计算个体匹配度并计算激活权重。但是由于缺失了部分信息，将导致推理的不准确性增大。

基于上述想法，本文提出了一种新的处理不完整数据集的BRB模型--基于扩展规则的析取范式置信规则库，该方法采用扩展置信规则库的方法，利用数据生成析取范式的规则库，并引入衰减因子对规则权重进行调整，当生成规则的数据含有缺失项越多，那么衰减因子便会调整规则权重，使其规则权重越小，若数据没有缺失，那么衰减因子为1。该方法能更加充分地利用缺失数据所含有的信息，并且在不增加系统复杂性的情况下，确保推理的准确性。

### 规则表示

基于EBRB快速构建规则库和析取范式的规则结构的特点，建立基于析取范式的EBRB系统，规则表示如公式(3-11)。该规则结构与EBRB一致，但采用””进行规则属性间的连接，并引入衰减因子，代表第k条规则的衰减程度，由规则中缺失属性权重和缺失个数所决定。

|  |  |  |
| --- | --- | --- |
|  |  | 公式(3-12) |

### 激活权重计算公式

由于规则形式进行改变，使得规则属性具有逻辑或的关系，因此，将公式(3-11)中得累乘符号进行替换，激活公式定义如下：

|  |  |  |
| --- | --- | --- |
|  |  | 公式(3-13) |
|  |  | 公式(3-14) |

其中，表示衰减因子，则为对规则权重进行调整后的值。

### 前提属性参考值个数的可变

本文提出的基于析取范式的扩展置信规则库，通过数据建立具有析取范式的置信规则库，既能避免繁复的参数训练过程，又能保留析取范式规则的特点。一般而言，BRB中前提属性参考值的个数可以由领域专家经验知识来设置，同时前提属性参考值的个数将决定规则库的规模，而Liu提出的EBRB中除了通过专家设置，还能直接对数据集中的数据均匀取个值作为前提属性参考值的效用。但对于本文提出的基于析取范式的EBRB，当前提属性参考值个数取值不同时，系统的推理准确率将会有明显的差异。以函数拟合实验为例，选取非线性函数：

|  |  |  |
| --- | --- | --- |
|  |  | 公式(3-14) |

选取五个值作为评价结果等级效用，规则前件属性为单属性，比较前提属性参考值个数取值不同时，基于析取范式的EBRB的推理性能表现情况。在区间内选取500个点作为数据集，用来生成规则库，再均匀的选取1000个点作为测试数据，用来比较系统的推理准确率。图3-1为训练数据集，图3-2为在不同的前提属性参考值个数上的拟合效果。

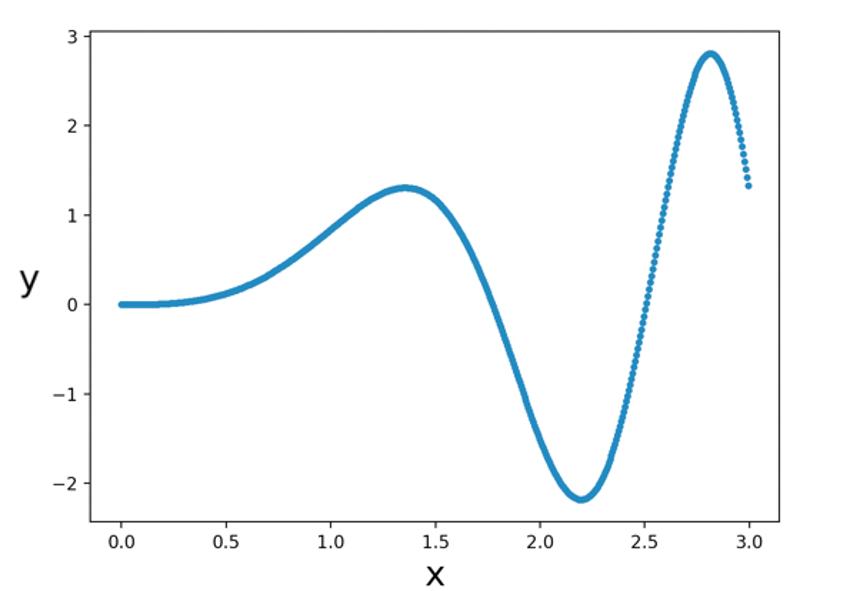


图3-1 函数拟合数据集

由图3-2可知当函数参考值个数取值较小时，基于扩展规则的DBRB只能对目标函数进行粗略的拟合，随着前提属性参考值个数的不断增加，可以看到拟合效果越来越好，最终与目标函数很好的重合。但是当前提属性参考值个数取值到达30时，基于扩展规则的DBRB已经能很好的对目标函数进行拟合，而后随着前提属性参考值个数的增加，并不会有更好的拟合结果。

表3-1记录了在不同的前提属性参考值个数下的平均激活规则数、平均推理时间和平均绝对误差。可以观察到当前提属性参考值个数为5时，平均激活规则数为227，激活率为45.7%，几乎是每一条输入数据都能激活近一半的规则，这将会造成过多不必要的规则被激活，被激活规则的评价结果置信度将会对最终的评价结果贡献一定的置信度，所以过多不必要的规则被激活将会造成最后推理结果准确率下降。随着前提属性参考值个数的增加，从表中可以看到函数拟合的平均绝对误差开始快速下降，在属性参考值个数取值30之后开始逐渐平缓。观察平均被激活的规则数，可以看到规则激活率从45.7%快速下降到22%，最后保持在3%左右，平均规则激活率下降的同时，平均绝对误差与之呈正比的趋势开始下降。从上述现象可以知道，当前提属性参考值个数增大时，被激活的规则数开始减少，每个输入数据可以更精确的激活对应的规则，减少多余的规则匹配可以减少冗余的信息，使得最后的决策结果更加准确。但是当前提属性参考值个数大于30时，发现平均绝对误差下降趋势开始趋于平缓，而平均推理时间仍然保持线性增加，因此，对于前提属性参考值个数的选择需要综合考虑平均绝对误差和平均推理时间。

表3-1 不同N值下的函数拟合效果

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| N | 规则数 | 平均激活规则数 | 平均规则激活率(%) | 平均推理时间(ms) | 平均绝对误差 |
| 5 | 500 | 227 | 45.7 | 5.899 | 0.450199 |
| 10 | 500 | 110 | 22.0 | 6.598 | 0.182863 |
| 15 | 500 | 72 | 14.4 | 7.995 | 0.105506 |
| 30 | 500 | 35 | 7.0 | 12.643 | 0.056142 |
| 50 | 500 | 21 | 4.2 | 19.145 | 0.046686 |
| 70 | 500 | 15 | 3.0 | 24.724 | 0.044714 |

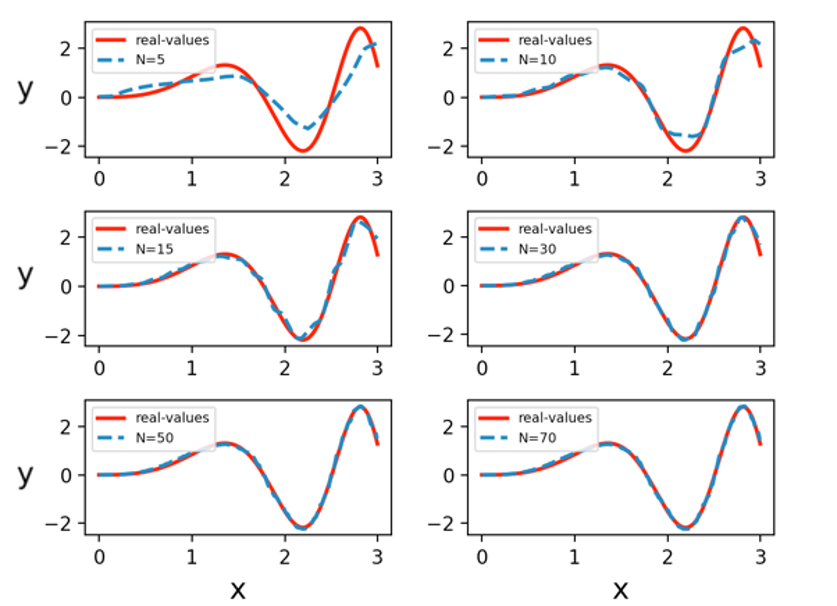


图3-2 在不同N下的函数拟合效果

### 对缺失数据的处理

基于析取范式的EBRB对不完整数据集处理的核心思想是：通过不完整数据生成不完整规则，并引入衰减因子对规则权重进行调整，通过缺失值的权重和个数对调整生成的不完整规则权重。假设输入数据，属性权重为，当第个属性值为缺失项时，那么基于析取范式的EBRB进行规则构建时，第个属性的置信分布情况是未知的，此时规则将处于不完整状态，按照缺失数据的属性权重，衰减因子的计算方式如下：

|  |  |  |
| --- | --- | --- |
|  |  | 公式(3-15) |

其中，表示规则中第个属性为缺失项。从公式可知，衰减因子与缺失的属性有关，当规则中没有缺失属性时，有；当规则中含有缺失属性时，有，当且仅当规则中的属性值完全缺失，此时经衰减因子调节后，规则权重为0，即该条规则在规则库中不再发挥任何作用。衰减因子与规则中缺失属性权重和缺失属性个数有关。基于析取范式的EBRB构建规则的步骤如下：

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效用；

步骤2：当数据不含有缺失项时，使用公式(2-4)计算数据前件后后件的置信分布形式，并采用第3.2.2节中的公式计算规则权重；

步骤3：当数据含有缺失项时，将缺失的属性的置信分布设置为0，表示该缺失的属性值无法匹配任何一个属性参考值；

步骤4：使用公式(3-15)计算衰减因子调整规则权重；

步骤5：输出带有基于析取范式的扩展置信规则；

基于析取范式的EBRB构建规则流程图如下：



图3-3 规则生成流程图

## 算法复杂度分析

本章节提出的基于析取范式的EBRB系统推理方法的时间复杂度主要体现在构建不完整规则库和推理部分。假设前提属性的个数为，每个前提属性具有个参考值，规则库规模为，评价结果等级总数为，那么生成每条规则的时间为，规则权重设置所需时间为，计算衰减因子所需要时间为，那么构建规则库总的需要时间为。本文方法对激活权重公式进行修改，而推理部分仍采用ER算法进行信息合成。将输入数据转化为置信分布形式需要，计算激活规则权重需要，对激活的规则进行ER推理，假设激活规则数为条，则ER推理所需要的时间为。因此，本文方法对每一条输入数据需要的时间为。

与Liu等[35]方法相比，本文方法在规则构建和推理部分具有相同的时间复杂度，但本文方法能够处理数据集不完整的情况，要优于EBRB方法。

## 实验分析

本节采用公共数据集中的Iris、Wine、Seeds、Transfusion、Mammographic五个数据集进行实验，表3-2对数据集信息进行说明。首先对比了在完整数据集下本文方法与其他EBRB方法的推理准确率，然后对了在不同缺失模式和缺失率下的推理准确率，验证本文方法的有效性。实验采用Intel [Corei5@1.4GHz;16GB](mailto:Corei5@1.4GHz;16GB)内存；macOS catalina操作系统；编程环境为python3.8。

表3-2数据集说明

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| No | 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 |

### 与其他EBRB的对比实验

#### 确定前提属性参考值个数

根据3.3.4节介绍看，前提属性参考值个数对基于析取范式的EBRB的推理性能有较大的影响。因此需要先确定各个数据集的前提属性参考值个数，以不同属性参考值个数下的分类准确率和推理时间作为评价依据。

在数据集上构建基于析取范式的EBRB时，前提属性参考值个数的取值分别为2，4，5，10，15，20个，以属性的最小值和最大值构成的区间内均匀取个作为前提属性参考值的效用。实验结果如图3-4所示。

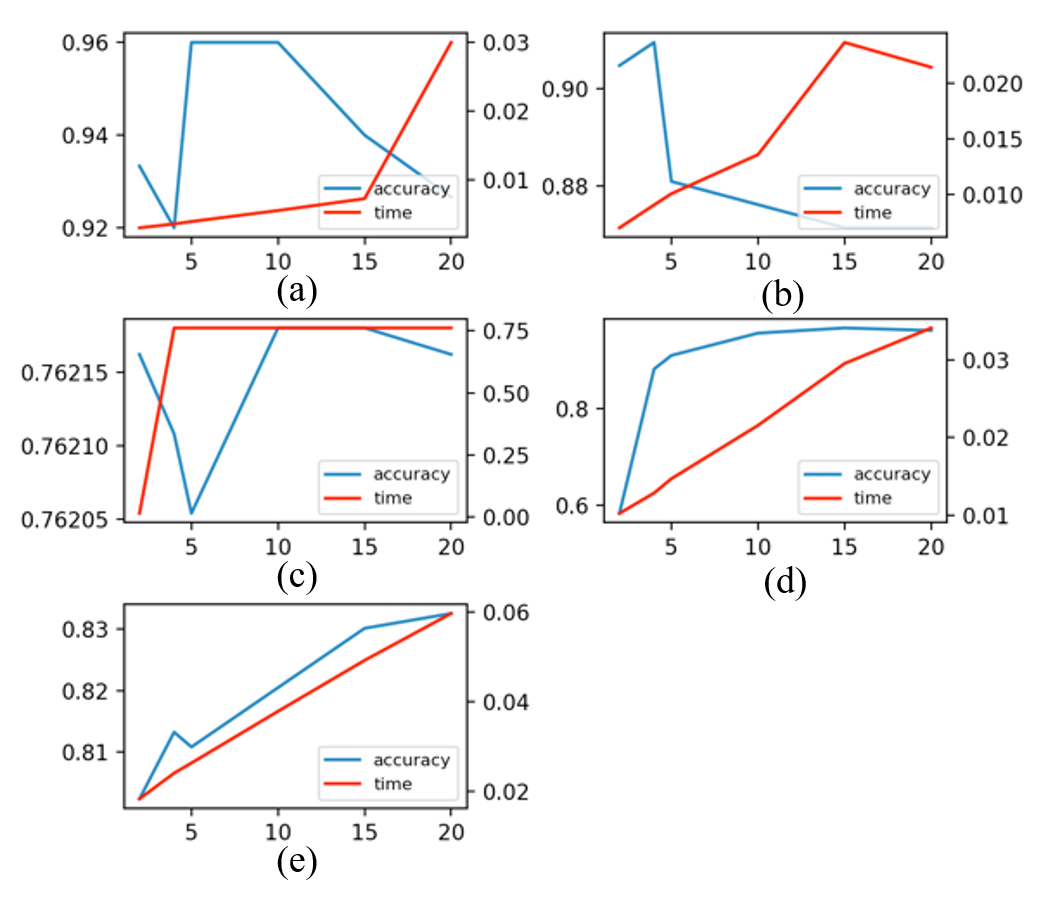


图3-4 不同参考值个数在五个数据集中的分类准确率：(a)iris数据集；(b)seeds数据集；(c)transfusion数据集；(d)wine数据集；(e)mammographic数据集

如图3-4所示，当N选取不同值时，不同的数据集具有不同的表现情况，对于iris数据集，总体呈现先升后降的趋势，当前提属性参考值个数取值为3到5时，分类准确率能达到96%，然而当个数超过10时，推理准确率反而开始下降，但整个过程中推理里所需要的时间保持增加，并以个数为15作为拐点，开始急剧增加，因此iris数据集最优的前提属性参考值个数取值为5；对于Seeds数据集，当参考值个数超过4时，分类准确率便快速下降，而推理所需要消耗的时间急剧增加，因此对于Seeds数据集选择4个前提属性参考值个数时，能达到最好的推理性能；对于Transfusion数据集，推理时间从开始便大幅度上升，但之后随着前提属性参考值个数的增加，推理时间变化较小，而分类准确率呈现先下降后上升的走势，最高准确率出现在附近，但观察纵坐标可知对于Transfusion数据集，前提属性参考值个数的变化，对最后的推理准确率影响较小，图3-5(c)中变化的幅度仅为0.2%，所以Transfusion的前提属性参考值个数取值为2较为合适；对于wine数据集，分类准确总体为上升趋势，但当时，分类准确增加变得平缓，同时推理时间却是线性增加，因此前提属性参考值个数取值为10时能达到较好的分类准确率并保持较低的推理时间；对于Mammographic数据集，分类准确率增加接近于线性增长，但是当属性参考值个数超过10时，分类准确率增加开始减少，因此参考值个数取值为10。

#### 对比实验

本节对比实验主要验证本文方法在完整数据集下仍能有较好的表现，主要对比的EBRB方法模型有Liu-EBRB[35]、DRA-EBRB[41]、MVP-EBRB和VP-EBRB[46]、CABRA-EBRB[47]和EBRB-C[48]。在对比实验中，对于本文提出的方法，根据上小节分析设置各个数据集的最优前属性参考值个数，其他EBRB模型默认前提属性参考值个数均取5。实验结果如表3-3所示。

表3-3 与其他EBRB进行对比

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | 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** | **96.00** | **90.95** | **96.63** | **76.22** | **83.01** | |

从表3-3中可以看到本文方法在Mammographic数据集的表现比其他几种EBRB模型来得更好，比表现最好的Liu-EBRB方法高出4.15%，同时，本文方法在在Iris和Wine数据集上与表现最好的CABRA-EBRB方法相持平，但本文方法还能够对不完整数据进行处理，因此本文方法明显优于其他EBRB方法，在Seeds数据集上本文方法表现不佳，但仍具有较好的分类准确率，在Transfusion数据集上，本文方法虽然达不到最好，但比Liu-EBRB和CABRA-EBRB具有更好的推理准确率。

本节实验表明，基于析取范式的EBRB推理方法在完整数据集上与其他的EBRB模型相比具有较好的推理精度，其中，本文方法在Mammographic数据集上表现最好，在Iris和Wine数据集上表现次于Mammographic数据集，在Transfusion和Seeds数据集上表现较为一般。下节对不完整数据集进行比较实验。

### 不完整数据集的对比实验

本节实验主要验证本文方法对于不完整的数据集的处理效果。数据集的不完整程度主要体现在三个方面：数据缺失程度、数据缺失模式和数据缺失分布[49]。数据集中数据项缺失的程度高低是衡量数据集不完整程度的主要指标，本节实验对完整数据集的部分数据进行丢弃，通过人为的方式构造不完整数据集。数据集的缺失模式主要分为两种：单一属性缺失和多属性缺失。单一属性缺失模式指缺失的数据集中在某个属性值当中，其他属性值不含有缺失项，多属性缺失模式指在该不完整数据集中，一条数据可能含有数个缺失项的属性，甚至可能属性数据全部缺失。数据缺失分布可以分为随机分布和非随机分布，随机分布主要是指缺失的数据项是随机产生的与外界因素没有关系，非随机分布是指由外界干扰所造成的数据项缺失，例如一张表格因洒了墨水使得纸上数据被覆盖，此时被覆盖的数据是具有连贯性，即缺失的数据是紧密相连的。

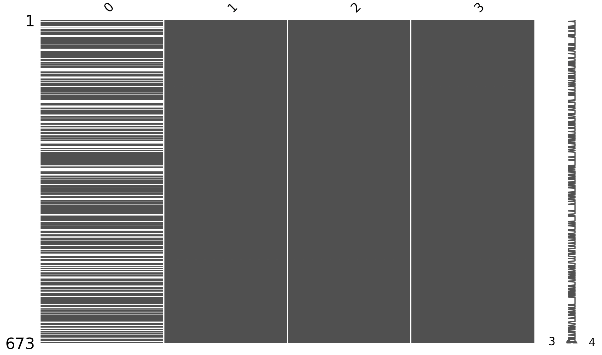


图3-5 第1个属性随机缺失30%的数据

本节实验主要采用单一属性属性随机缺失模式和多属性随机缺失模式，数据缺失程度依次取10%到90%。单一属性随机缺失模式的缺失项主要分布在某个属性中，缺失项的位置随机给定，例如图3-5，图中每一行代表一条输入数据，每一列代表一个属性，该数据集有673条数据，每条数据含有4个属性分别编号为0、1、2、3，图中黑线代表该数据未丢失，白线代表该数据缺失，右侧为统计数据集中最少和最多缺失的属性值。从图可知，数据集在编号0的属性上缺失了30%的数据，此时，该数据集含有完整的数据也含有至少3个属性未缺失的数据。

多属性随机缺失模式的缺失项不再限定于某个属性，而是在该数据的所有属性中可能会出现数个属性具有缺失值，其中缺失的属性是随机选择的，例如图3-6所示，673条数据中含有10%的缺失项，这些缺失项分布于数据集的各个地方，从右侧统计结果可以看到，该数据集中有四个属性都未缺失的数据，即完整的数据，也有缺失了三个属性，只剩下一个未缺失的数据。在更极端的情况下，可能会出现全部属性都缺失的情况。

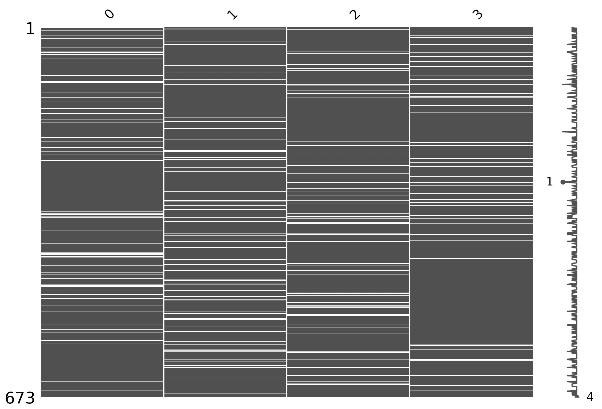


图3-6 多属性随机缺失10%数据

#### 与Liu-EBRB对比实验

Liu等[35]提出了EBRB方法，其他EBRB模型都是在Liu-EBRB的基础进行改进得到的，对EBRB的改进方面主要有个体匹配度公式优化、规则检索优化等，但这些改进的方法未考虑到数据缺失的情况，为了降低系统的复杂性，本节实验只和原始的Liu-EBRB进行对比实验。所对比的数据集仍采用Iris、Wine、Seeds、Transfusion和Mammographic五个数据集，其中通过两种缺失模式模式构建不完整规则库，数据集的缺失程度分别为10%到90%。由于EBRB无法直接处理具有缺失项的数据集，因此，本文采用均值、中位数和众数对缺失项进行填充。因为EBRB采用数据生成规则，当数据量不足时容易产生零激活问题，在本节实验中，对于出现的零激活输入数据，将其默认为分类错误，以便比较方法的推理准确性。本次实验采用20次十则交叉验证取总平均值进行比较。

图3-7为单属性随机缺失模式下的对比结果，从图可知本文方法对于不同缺失率下的数据集具有良好的平稳性。在Wine和Transfusion数据集中本文方法明显优于使用统计方法填充，并且随着数据的缺失率的增加，本文方法的推理准确率下降平缓，而使用填充方法的Liu-EBRB推理准确率下降幅度开始变大；在Iris、Seeds、Transfusion数据集中，本文方法在低缺失率下推理准确率比传统的统计填充来的略低，但随着数据的缺失率升高，本文方法的推理准确率开始出现略微的下降，而由统计填充的Liu-EBRB算法开始快速地下降。在Iris数据集中，缺失率为50%时本文方法与统计填充方法具有相同地准确率，而缺失率大于50%时，统计填充的方法开始快速下滑。本文方法的准确率开始大于统计填充方法，同样的在Transfusion数据集中，当缺失率大于70%时，统计方法的准确率比本文方法来得小，而在Seeds数据集中使用中位数填充的方法表现最好，且众数和均值填充的方法只在较高的缺失率下表现较差。

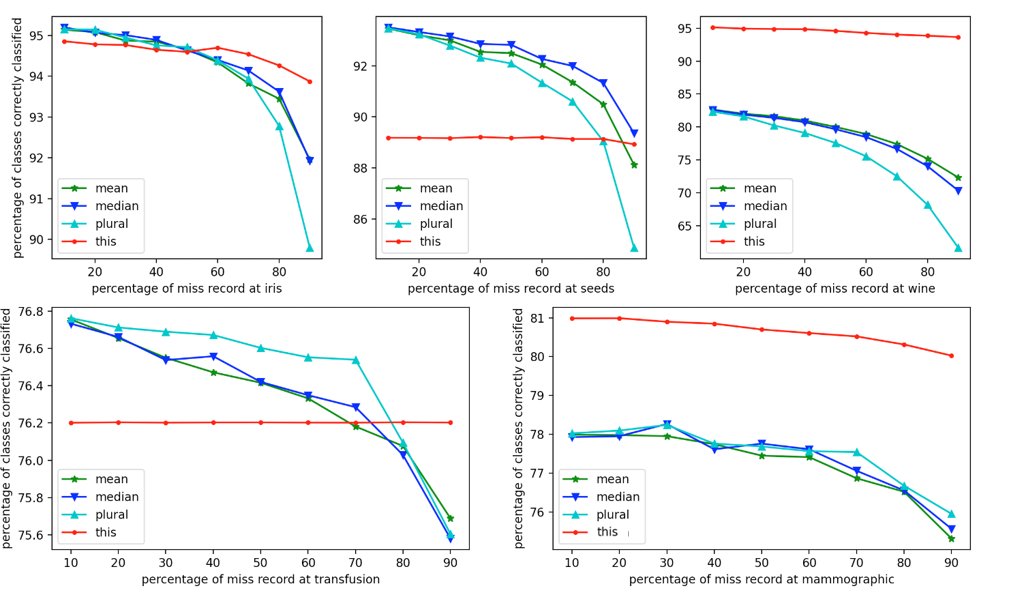


图3-7 EBRB单属性随机缺失实验

图3-8为多属性随机缺失模式下的对比结果，相比于单属性随机缺失，本文方法在较低的数据缺失率能与统计填充的方法具有相近的准确率。同样的，本文方法在wine和Transfusion数据集中推理准确率比统计填充方法来得更好，而在Iris、Wine和Transfusion数据集中，当数据缺失率大于一定值的时候，本文方法仍能够保持推理的准确并高于统计填充的方法。

无论在单属性随机缺失或多属性随机缺失，统计填充方法在Wine和Mammgraphic数据集上都表现较差，其主要原因在于采用统计方法进行填充时，填充的数据相对比较单一，例如中位数代表历史数据的“中间水平”，均值代表“平均水平”，众数代表“次数最多水平”，使用这些数据填充缺失项，将导致填补后完整的数据集中的数据缺失多样性，即多条数据将具有相同的属性值，当数据集缺失程度较大时，输入数据更容易引起“零激活”，使得最后推理准确性下降。在Transfusion数据集中，无论缺失程度如何，本文方法的推理准确率能保持在一个很小的波动，当缺失率较高时，容易出现属性值全部缺失的情况，此时该数据生成规则对应的衰减因子为0，那么该规则实际上并不会被激活。因此在高缺失率下的基于析取范式的EBRB中的规则库会比Liu-EBRB少，从而避免了激活不必要的规则。并且析取范式的规则结构更容易激活，不会造成大量的零激活输入。

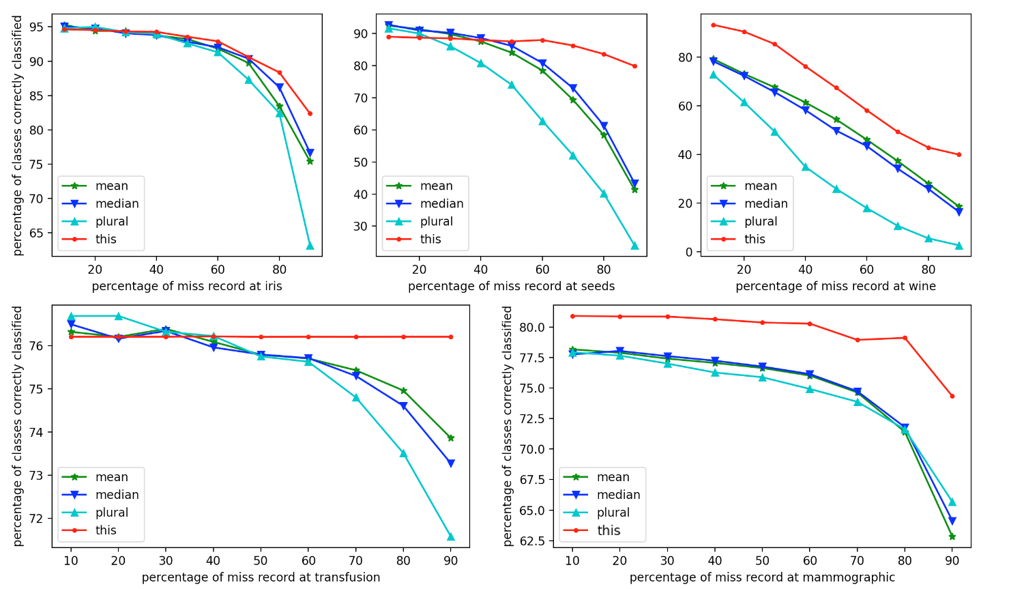


图3-8 EBRB多属性随机缺失实验

#### 与DBRB对比实验

本文方法根据DBRB析取范式规则的特点提出基于扩展规则的DBRB推理方法，本文方法即拥有DBRB的特点同时也克服DBRB面对缺失数据的缺点，本节内容主要与DBRB在不同缺失程度和缺失模式下进行对比。数据集的缺失模式和缺失程度仍与3.4.2.1节相同。

图3-9为在单属性随机缺失模式下的对比结果，图中明显本文方法的推理准确率高于DBRB方法。DBRB方法推理准确率会随着数据缺失程度的增大而大幅下降，这是因为缺失项的增加导致在DBRB参数训练时无法正确的设置权重等参数。而本文方法利用数据直接生成析取范式的规则，避免了不完整数据在参数训练时造成的影响。

图3-10为多属性随机缺失模式下的对比结果，本文方法依旧好于DBRB方法，当缺失程度较大时，本文方法会构建大量不完整规则，并通过衰减因子调整规则，抛弃完全没用的数据，这些不完整的规则被激活后将参与ER合成，未缺失项能够对最终的评价结果置信度提供一定的支持。因此，本文方法一方面避免了不完整数据集的参数训练，另一方面通过衰减因子对不完整规则进行调整，使该不完整规则为推理结果提供合理的支持。

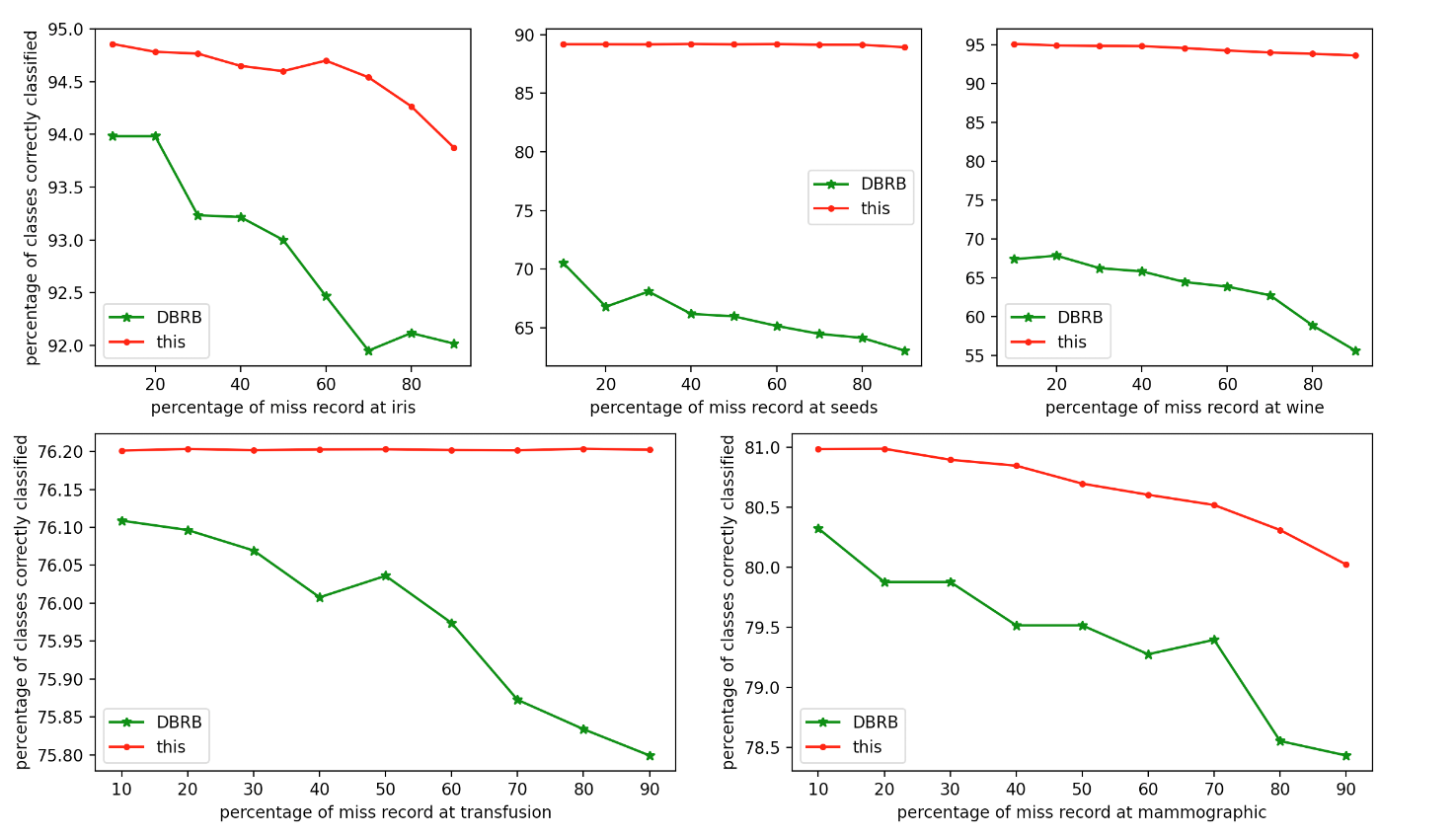


图3-9 DBRB单属性随机缺失实验

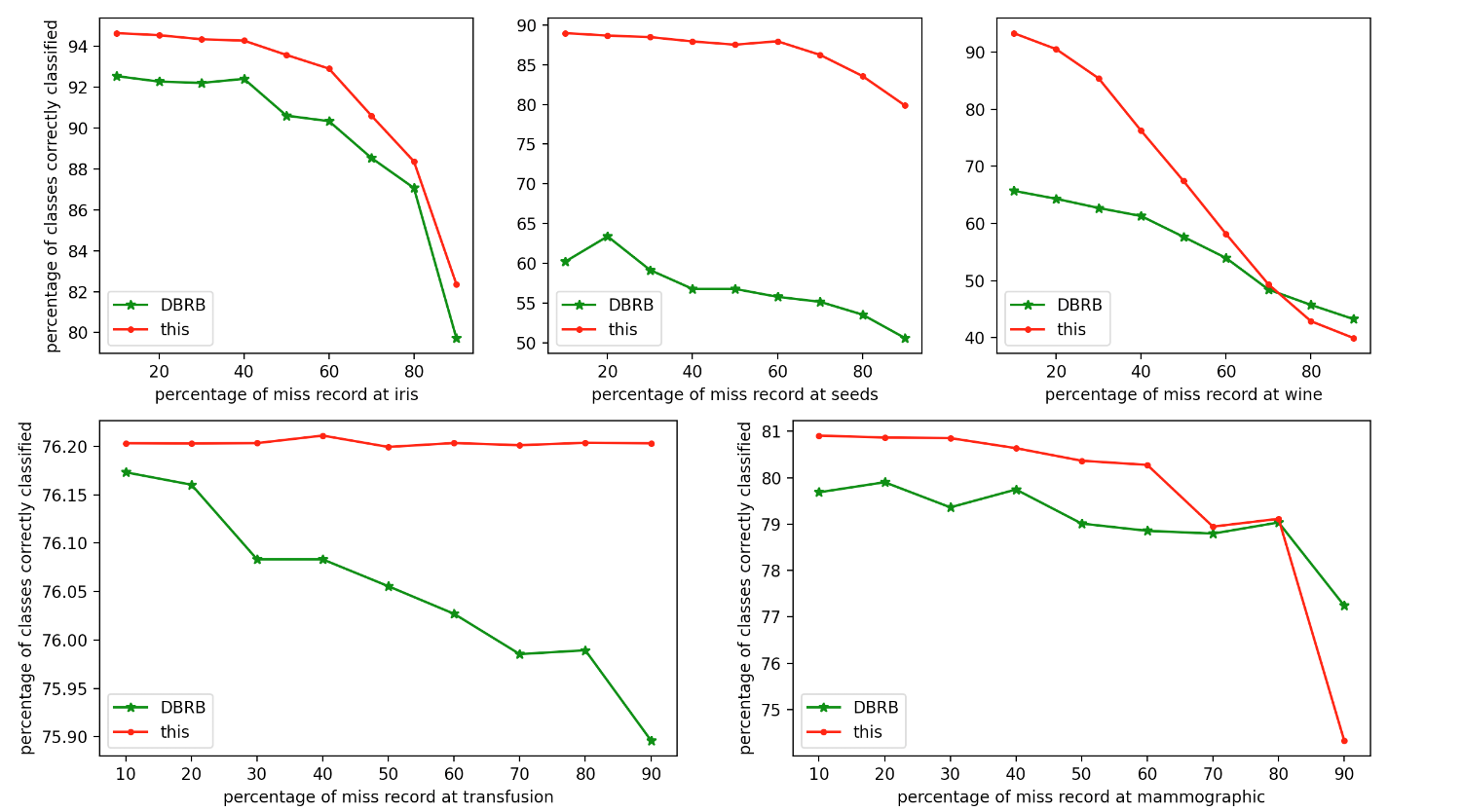


图3-10 DBRB多属性随机缺失实验

本节实验结果表明，本文方法能有效处理不完整数据集。在与EBRB的对比实验中，本文方法能有效处理类别分布不明显的不完整数据集，在低数据缺失率中保持较好的推理准确率，在高缺失率中具有很好的推理性能。在与DBRB的对比实验中，本文方法对缺失数据的处理明显优于DBRB方法，同时避免了参数训练带来的性能消耗。

## 本章小结

本章针对BRB中不完整数据集问题，根据EBRB和DBRB无法处理缺失数据的缺点，提出了基于扩展规则的DBRB方法。该方法能够利用历史数据建立带有析取范式的置信规则库，同时当数据集时不完整时，该方法能够相应的建立不完整的规则。此外，本章还引入了衰减因子用来对规则权重进行调节，若数据还有缺失项时，可以通过缺失属性权重和缺失属性个数和属性权重来计算衰减因子，通过衰减因子对规则权重进行调节，当数据缺失的属性越重要时或缺失项的个数很多时，规则权重会相应的变小。在最坏情况下，该条规则权重将为0；若数据没有缺失项时，那么本章方法的计算方式与EBRB基本一致。因此，本章方法能够在不增加系统复杂度的情况下，进行不完整数据集地处理。为了验证本文方法，使用公共测试数据集进行实验，首先与其他EBRB模型进行对比，证明本章方法在完整数据集中能与其他EBRB优化模型取到较好的推理准确率，然后与EBRB和DBRB进行不完整数据集处理的对比试验，证明本章方法能有效的对不完整数据集进行推理，并保证一定的推理精度，研究结果表明本章所提的研究方法取得了预期的效果。

# 第四章 基于模糊关联规则的DBRB优化方法

## 引言

在第三章提出了基于扩展规则的析取范式置信规则库优化方法，该方法能有效地处理不完整数据集，完成规则库的构建与推理。同时，DBRB系统是一个能够处理带有模糊性不确定性和概率不确定性信息的决策系统，它能够对用户输入的数据进行推理并做出相应的决策。但当用户输入的信息是不完整时，DBRB系统难以将输入信息与规则库中的规则进行匹配，从而导致系统难以进行正确地做出决策。因此，如何处理不完整信息决策问题，对DBRB系统具有重要的实际意义。

目前相关研究有，Yu等[50]通过采样思想，通过历史数据估计样本分布并对样本进行分层抽样填补缺失的数据，对多个缺失填补得到的数据进行ER推理生成填补数据的合成结果，然后再次使用ER算法对各个合成结果再次融合得到最终的输出，该方法朴素易懂但实际数据集中含有噪声或离群点会使得该方法无法正确估计样本分布；Li等[51]利用BRB系统处理不准确性、不确定性数据的能力，对观测到的不完整数据进行ER合成缺失的部分数据，最终得到一个完整观察结果，再进行BRB推理，该方法成功应用到医疗评估中，但该方法只适用于单个属性缺失的情况，无法处理含有多个缺失的输入。

针对上述问题，本章方法针对不完整的输入信息提出基于模糊关联规则的析取范式置信规则库的推理方法，该方法可以对缺失的输入数据进行模糊化，通过关联规则推断缺失数据的置信分布并填充，然后将该填充后的置信分布形式的数据进行DBRB进行推理。为了验证本文方法的有效性，采用公共测试数据集中的Breast Cancer、Iris和Transfusion数据集进行测试，并与其他方法进行对比。

## 构建模糊置信表示库及DBRB规则库

模糊集由Zadeh等[52]在1965年提出，随后在此基础上模糊理论[6]开始快速发展，模糊理论允许将集合的隶属度表示为一种可能性分布，该分布取决于该集合的置信程度而不是概率。通过模糊化语义表述数据，我们可以使用隶属度函数计算每个数据所属集合的隶属度从而得到一种数据的模糊化表示方法。本节内容主要分为两个部分：构建模糊置信状态库和构建DBRB规则库。其中，利用隶属函数构建模糊置信状态库，同时，数据也将通过参数训练的方式用于构建DBRB规则库。

### 前提属性参考值的模糊表示

在Yang等[9]提出的RIMER方法既可以定量输入数据也可以处理定性的输入数据。对于定量的数据可以采用基于规则

和效用的信息转化方法获取输入数据的表示，计算方式为公式(2-4)。该计算公式可以视为三角隶属函数，例如，假设输入数据为，决策系统中第个属性的参考值效用为，那么第个属性参考值的隶属度函数如下：

|  |  |  |
| --- | --- | --- |
|  |  | 公式(4-1) |

对于第个属性中参考值的隶属度分布如图4-1(a)所示。从图可知，输入数据的第个属性和第个属性参考值的隶属度由与的距离所决定，隶属度呈线性增加。然而这种表示方法是静态且无法满足不同数据属性特征。在隶属度计算中，不同的数据集的属性参考值并不一定呈线性的关系。由二八定律启发，让仅有20%的数据操控80%的结果。假设对于越靠近目标效用值的输入数据具有越高的隶属度，那么该输入便越能反应其数据特征，反之当输入数据离目标效用较远时应该具有较小的隶属度，较小该数据在该特征上的影响。因此，对于某个参考值的隶属度应当是集中于某一值附近，而不是呈线性关系。

为了解决上述问题，本文使用高斯隶属函数替换原有的三角函数计算，高斯隶属函数如图4-1(b)。公式如下：

|  |  |  |
| --- | --- | --- |
|  |  | 公式(4-2) |

其中，取值越小，函数越集中于附近。，当且仅当。通过使用高斯隶属度函数，可以将输入数据的隶属度集中在属性参考值附近，同时可以针对不同的数据集修改对应的，调整隶属度的分布情况。

由公式(4-2)可知，假设输入数据的第个属性位于和之间，对应的隶属度为和，那么可能并不等于1，因此，需要对其归一化：

|  |  |  |
| --- | --- | --- |
|  |  | 公式(4-3) |

因此，可以通过公式(4-3)计算输入数据的隶属度和DBRB中的规则个体匹配度。

### 构建模糊置信表示库

模糊置信规则由各个属性的隶属度构成，对于单个数据可以表示为，其中，表示第个属性第个参考值，表示隶属于第个属性第个参考值的程度。模糊置信规则库可以表示为。假设第个属性为感冒程度，它有3个参考值分别为{轻微，中等，严重}，通过隶属度公式计算可得到，表示该数据隶属于轻微集合的程度为0.4，隶属于中等的程度为0.6。通过对每个属性使用公式(4-3)，那么可以得到该条数据的模糊置信表示，对数据集中的所有数据进行相同的操作，便可以得到该数据集的模糊置信表示库，如表4-1所示。

表4-1模糊置信表示库

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | | |  |  | | | … |  |  | | | D |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  | 0.4 | 0.6 | 0 |  | 0.1 | 0.9 | 0 | … |  | 0.2 | 0.8 | 0 |  |
|  | 0 | 0.2 | 0.8 |  | 0 | 0.3 | 0.7 | … |  | 0 | 0.4 | 0.6 |  |
| ... | ... | ... | ... |  | ... | ... | ... | ... |  | ... | ... | ... | ... |
|  | 0 | 0.5 | 0.5 |  | 0 | 1 | 0 | … |  | 0.5 | 0.5 | 0 |  |

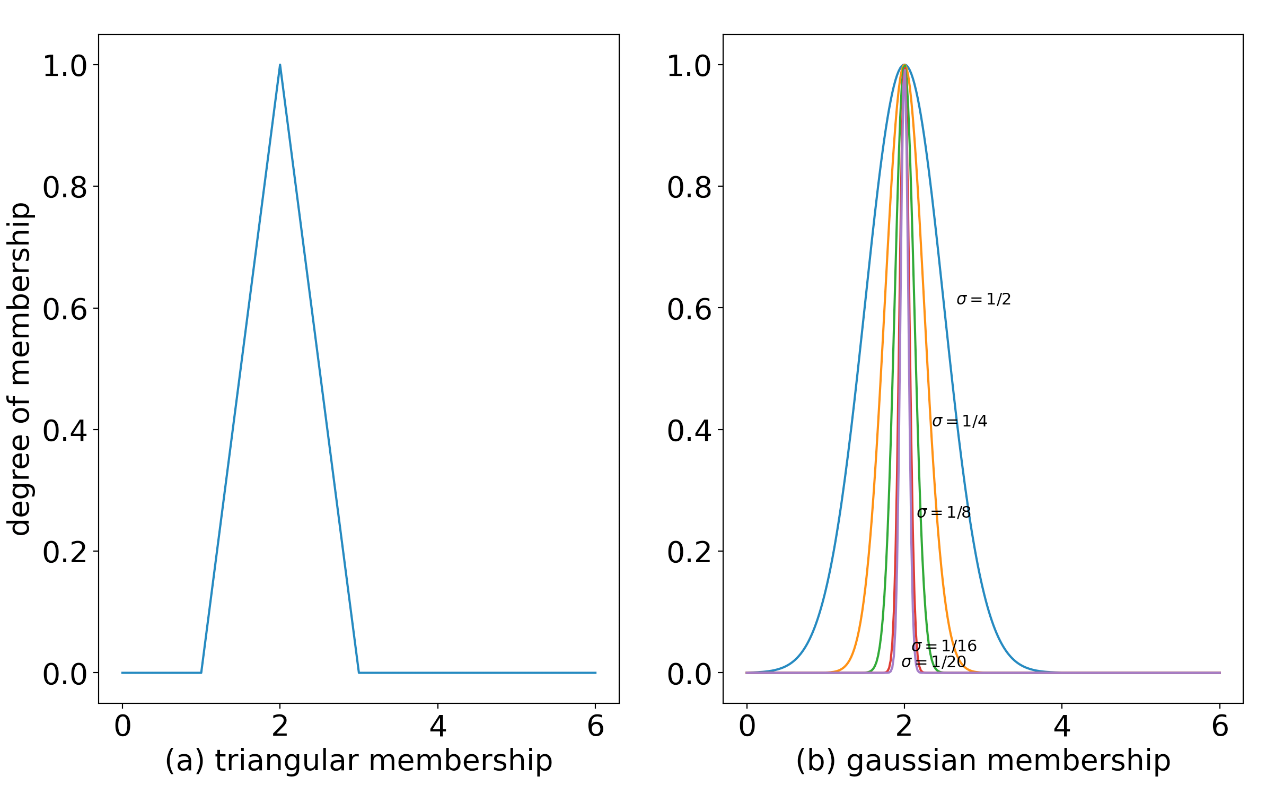


图4-1属性参考值隶属度

### 基于差分进化的DBRB

Yang等[9]提出的DBRB采用加权乘积和聚合函数组合个体匹配度并计算规则激活权重，Zhang等[53]对该公式进行重新定义，提出该计算公式符合容斥原理，使该计算方式具有更好的数学含义。因此，本节使用的DBRB激活权重采用Yang等[16]所提出的计算公式。

根据领域专家知识经验初始化DBRB参数，其中前提属性参考值，可以通过构成的区间均匀取5个点，将这五个值作为属性参考值的效用。然而专家知识的储量将会影响到参数设置的好坏，因此为了更好的提高DBRB推理性能，需要进行参数训练。Chang等[19]提出DBRB参数训练方法，成功应用于分类问题，Yang等[21]人在Chang的基础上提出了动态参数训练方法，在桥梁风险评估进行应用。这样方法虽然取得较好的效果，但在多属性多参考值的问题中增加了需要训练的参数个数，增加了在参数学习上的时间花销。因此，本节方法的参数训练模型将不考虑属性参考值效用，只对规则属性权重、规则权重和规则评价结果等价置信度进行参数学习，该方法能在少量的训练参数的情况下保证系统的推理准确性。训练模型的目标函数如下:

|  |  |  |
| --- | --- | --- |
|  |  | 公式(4-4) |

其中，为DBRB输出结果与实际结果的均方根误差。该模型含有如下几个约束条件：

1. 第条规则的第个属性的第个参考值需要满足以下限制：

|  |  |  |
| --- | --- | --- |
|  |  | 公式(4-5) |

1. 第条规则权重需要满足以下限制：

|  |  |  |
| --- | --- | --- |
|  |  | 公式(4-6) |

1. 第条规则的评价结果等级需要满足以下限制：

|  |  |  |
| --- | --- | --- |
|  |  | 公式(4-7) |

当第k条规则是信息完全的时候，有，否则，；

差分进化(Differential Evolution，DE)算法是一种群智能算法，最早由price等[54]提出。该算法相比于其他的群智能算法具有更高的准确率和更快的收敛速度，同时该算法简单易实现且高效，已经被广泛地应用于参数训练当中。DE算法伪代码如算法4-1所示。

本节首先提出了基于高斯隶属函数的个体匹配度计算公式，该方法针对传统的个体匹配度计算方式的不足进行改进，然后介绍了模糊置信表达库和DBRB规则库的构建方法。模糊置信表达库是本章方法的基础，模糊置信表达库是通过隶属度函数表示数据对于各个属性参考值的隶属程度，通过该步骤对输入数据进行模糊化表示，使输入数据能更好地表示所含有的信息。同时，输入数据还将被用来构建DBRB置信规则库，通过求解目标函数最优值获取适合于DBRB的参数设置，训练算法采用了DE算法对训练模型进行求解，该算法能快速高效地求解最优问题，使DBRB规则库构建更能加高效。本节的步骤流程图如图4-2所示。

|  |
| --- |
| 算法4-1：差分进化算法  输入：训练集X，种群规模NIND，最大遗传代数  输出：规则属性权重，规则权重，规则评价结果置信度，参考属性效用； |
| 1. Pop = initialization() //初始化种群 2. **for** i =1 **to** MAXGEN: 3. **for** j =1 **to** NIND: 4. mutation(pop[j]); //变异操作 5. u = crossover(pop[j]); //交叉操作 6. dbrb = buildDBRB(u); //构建DBRB 7. mse = MSE(dbrb, X);//计算均方根误差 8. selection(u, pop[j], mcf)//选择操作 9. **return** bestFit(pop);//返回最优适配度的个体 |

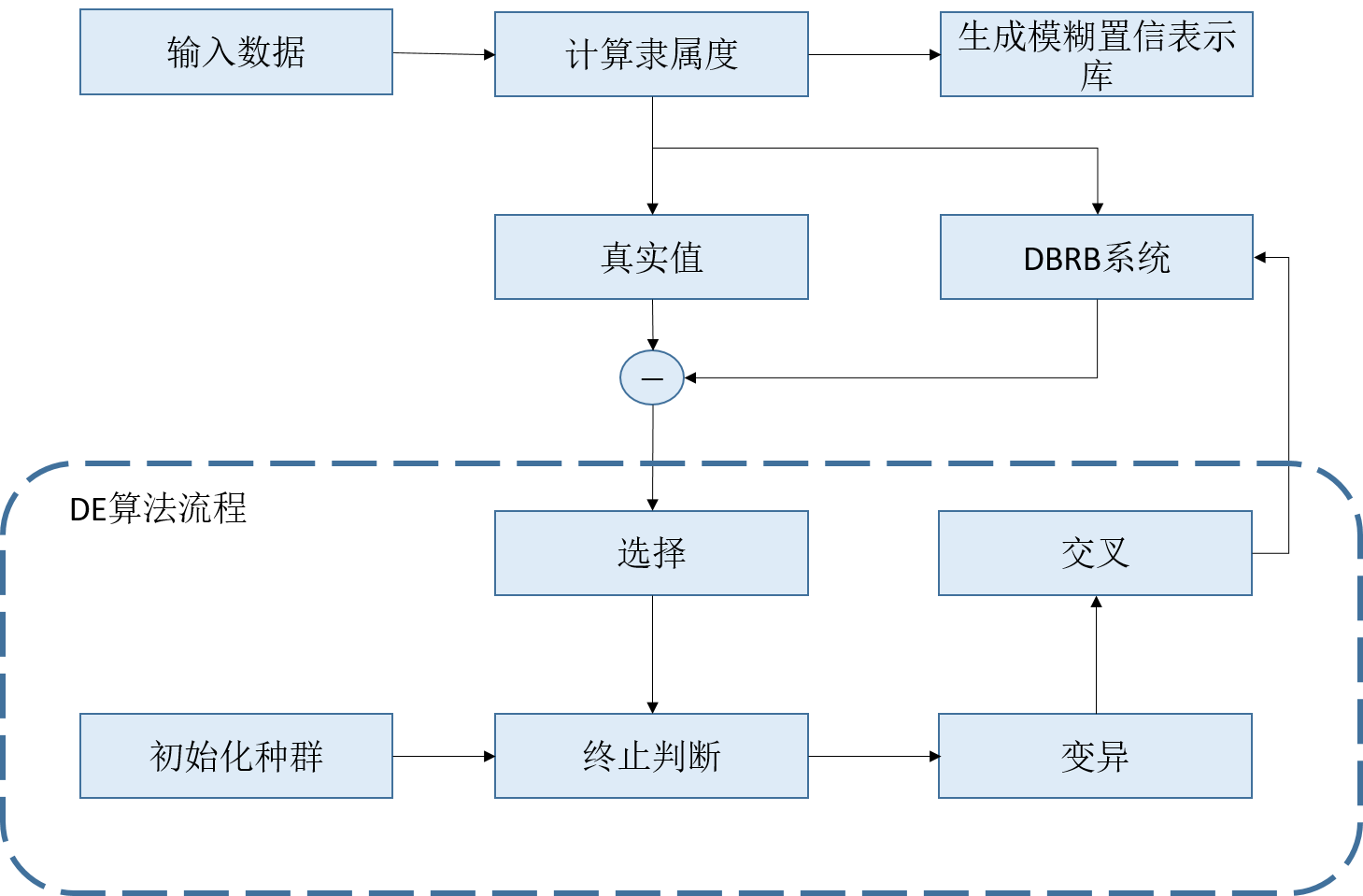


图4-2构建流程

## 关联规则推断缺失值方法

### 关联规则

通过上节步骤，我们可以获得模糊置信表示库，其本质上是存储各个输入数据对属性各个参考值的隶属程度，当用户输入的信息含有不完整数据时，那么可以通过该模糊置信表示库构建关联规则，进而推断缺失数据的置信分布表示形式。假设有两个属性集分别为和且，则关联规则可以表示为：

|  |  |  |
| --- | --- | --- |
|  |  | 公式(4-8) |

其中，为，为，为属性值的语义表示，符号表示推断，即从可以推断出。例如，根据隶属度为的“高”和隶属度为的“低”，可以推断出属性为“高”的隶属度为。通过这种关联规则形式，我们根据模糊置信表示库构建多个关联规则。如表4-2中带波浪线数据，在中有。将若干条具有相似结构的关联规则进行规则组合能够得到对应属性参考值的隶属度，即 ：

|  |  |  |
| --- | --- | --- |
|  |  | 公式(4-9) |
|  |  | 公式(4-10) |

其中，为组合的关联规则的均值，表示总关联规则的个数，表示第k条关联规则对第i个属性第j个参考值的推断结果。例如表4-2中带下划线数据，有，因为在中只有和的和的隶属度不为零，因此。该关联规则表示第a个属性对第b个属性第j个参考值的推断结果。

表4-2 关联规则示意表

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | | |  |  | | | … |  | | | D |
|  |  |  |  |  |  |  |  |  |  |  |  |  |
|  | **0.4** | **0.6** | 0 |  | **0.1** | 0.9 | 0 | … | 0.2 | 0.8 | 0 |  |
|  | 0 | **0.2** | **0.8** |  | 0 | **0.3** | 0.7 | … | 0 | 0.4 | 0.6 |  |
| ... | ... | ... | ... |  | ... | ... | ... | … | ... | ... | ... | ... |
|  | 0 | **0.5** | **0.5** |  | 0 | **1** | 0 | … | 0.5 | 0.5 | 0 |  |

通过上述可知，通过模糊置信表示库可以构建关联规则，并将多个具有相同前件结构的关联规则进行计算，得到某个属性的隶属度。但是在计算过程中，我们并不考虑推断结果为0的关联规则，即为的关联规则，这是因为该条规则对推断结果不提供任何支持。但完全不考虑该规则，则会导致部分信息丢失，因此，引入一个打折系数：

|  |  |  |
| --- | --- | --- |
|  |  | 公式(4-11) |

其中，分子表示关联规则中结果的隶属度不为0的规则个数，分母表示关联规则中满足关联规则结构的总个数。此时，RS规则将表示如下：

|  |  |  |
| --- | --- | --- |
|  |  | 公式(4-12) |

例如表4-2中，有，因为在和中只有的推断结果的隶属度不为0，所以，而。

### 缺失数据的推断方法

当输入数据含有缺失部分时，可以通过从模糊置信表示库中构建关联规则，并利用相关规则中的信息获取缺失数据的置信表示形式。假设输入数据中的第2个属性值为缺失项，首先计算未缺失属性参考值的模糊隶属度，如表4-3所示；然后根据缺失项构建关联规则，例如为缺失项，那么可以结合第一个属性参考值构建一条关联规则为；再从模糊置信表示库中寻找具有相同结构的关联规则，在表4-2的第一属性只有的和可以推断出；对所有属性搜寻到的关联规则，有：

|  |  |  |
| --- | --- | --- |
|  |  | 公式(4-13) |

其中，表示第个属性中关联规则的个数，表示所有规则的个数。那么，最终可以得到缺失属性值的隶属度分布为。

表4-3缺失数据的模糊置信表示

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | | |  |  | | | … |  | | | D |
|  |  |  |  |  |  |  |  | … |  |  |  |  |
|  | 0.7 | 0.3 | 0 |  | ? | ? | ? | … | 0.2 | 0.8 | 0 |  |

通过模糊关联规则获取到缺失数据的隶属度分布情况后，便可以将该条数据作为DBRB输入数据进行决策推理。总体推断如算法4-2所示。

本节介绍了基于模糊关联规则的DBRB优化方法，在DBRB系统中添加模糊置信表示库，使得训练数据集不仅对DBRB系统进行参数训练，同时还保存数据的模糊表示信息，对不完整的输入数据模糊化，然后进行关联规则推理缺失部分的隶属度分布并填补得到完整的数据，最后使用DBRB对输入数据的隶属度分布进行推理。本节方法流程如下：

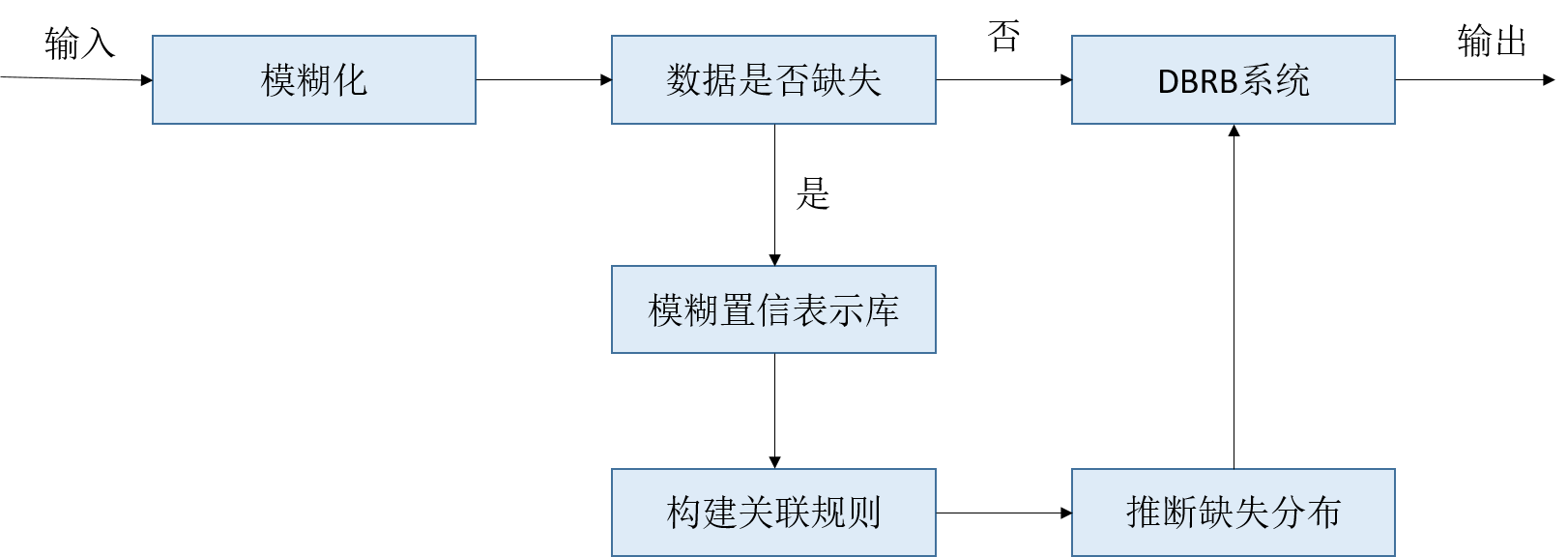


图4-3基于关联规则的DBRB系统流程

## 算法复杂度分析

本章提出基于关联规则的DBRB优化方法的时间复杂度主要分为规则库构建和推理两个部分。假设输入数据个数为，每个数据有个属性，每个属性有个参考值，每条规则有M个评价结果指标。规则库构建时，输入的完整数据将通过隶属度函数建立模糊置信表示库，时间复杂度为，计算个体匹配度所需时间为，计算激活权重时间为，使用ER推理所需时间为，参数训练由所使用的DE算法的迭代次数和每次迭代时间复杂度所决定，所以构建规则所需要的时间为。

进行不完整信息决策时，先计算完整数据的隶属度分布，时间复杂度为，然后对缺失项搜寻关联规则，时间复杂度为，然后是计算激活权重并推理，时间复杂度为，决策所需要的时间为。

从上述分析可知，本文方法在构建规则时需要构建一个模糊置信表示库，但不会增加构建规则的时间复杂度，而进行不完整信息决策时需要额外消耗时间搜寻关联规则。

|  |
| --- |
| 算法4-2：数据缺失推断算法  输入：第个属性缺失的不完整信息数据、模糊置信表示库  输出：完整数据的隶属度分布 |
| 1. 计算非缺失项数据在各个属性参考值中的隶属度； 2. **for** **to** : 3. 初始关联规则队列、空规则队列和结果队列; 4. **for** l = 1 **to** T: 5. **for** k = 1 **to** L: 6. **if** 与具有相同的结构 **then**: 7. **if** **then**: 8. ; 9. **else** **then**: 10. ; 11. **end** **if** 12. **end if** 13. 使用公式(4-9)和(4-10)计算和，并; 14. **end for** 15. 使用公式(4-13)计算缺失属性参考值的隶属度，更新； 16. **end for** 17. **return** ； |

## 实验分析

为了验证本章方法的有效性，选取公共测试数据集中的数据集进行测试。首先验证本文方法对不完整信息决策的有效性，然后与采用其他填补数据方法的DBRB进行比较，最后将本文方法与其他方法进行对比。实验环境为Intel(R) Core [i7-6700@3.40FHZ；16GB](mailto:i7-6700@3.40FHZ；16GB)内存；windows10操作系统；编程环境为Python3.8。

### 方法验证实验

本节使用Breast Cancer数据集对基于模糊关联规则的DBRB推理方法进行测试，分别用三组实验，证明当输入数据不完整时，本文方法能够得到较好的不完整信息决策结果。Breast Cancer数据集有9个属性和两个类别标签，该数据集共有699条数据，其中16条缺失第6个属性的数据，所以该数据集含有683条完整数据集。对DBRB设置前提属性参考值，依次在每个属性构成的最大最小值区间中，均匀取五个点作为该属性的五个参考值效用，DE算法设置初始种群个数为100个，最大迭代次数为100次。

首先使用683条完整的数据对DBRB进行参数训练，然后使用16条不完整的数据作为测试集进行输入，得到输出结果如表4-4所示，其中编号为563649的数据错误分类，其余标签均正确地标识，准确率为93.75%。

表4-4 breast cancer数据集不完整数据测试结果

|  |  |  |
| --- | --- | --- |
| 属性编号 | 输出标签 | 实际标签 |
| 1057013 | Malignant | Malignant |
| 1096800 | Benign | Benign |
| 1183246 | Benign | Benign |
| 1184840 | Benign | Benign |
| 1193683 | Benign | Benign |
| 1197510 | Benign | Benign |
| 1241232 | Benign | Benign |
| 169356 | Benign | Benign |
| 432809 | Benign | Benign |
| **563649** | **Benign** | **Malignant** |
| 606140 | Benign | Benign |
| 61634 | Benign | Benign |
| 704168 | Benign | Benign |
| 733639 | Benign | Benign |
| 1238464 | Benign | Benign |
| 1057067 | Benign | Benign |

对于完整的数据，使用十次十折交叉进行测试，对测试数据集依次丢弃每个属性的数据，使得数据集中始终有一列属性值是缺失项，遍历所有属性便可以得到，当缺失各个属性值时，本文方法的推理准确率。实验结果如表4-5所示，表格中第i个属性表示该属性值全部缺失，而第i属性值之外的数据没有缺失，从表中可以看到，本方法在输入数据为单个属性缺失的情况下仍能够达到较好的分类准确率，最高平均分类准确率为99.51%，最低平均为93.95%，总平均分类准确率为96%。

表4-5 十次十折单属性缺失

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 缺失属性 | | | | | | | | | 平均值 |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| min(%) | 92.65 | 94.12 | 94.12 | 94.12 | 94.12 | 94.12 | 94.12 | 94.12 | 94.12 | 93.95 |
| max(%) | 98.55 | 100 | 100 | 100 | 100 | 100 | 100 | 97.10 | 100 | 99.51 |
| avg(%) | 94.71 | 95.89 | 96.18 | 96.33 | 96.04 | 96.33 | 96.18 | 95.89 | 96.48 | 96.00 |

下面考虑在任意多个输入属性具有缺失的情况，选取数据集中的(0,3,8),(1,4,6),(2,5,7)属性作为缺失值，实验结果如表4-6所示。当数据中第1，4，6个属性缺失时，推理的平均准确率仍可以达到97.22%，而第2，5，7个属性缺失时，此时填补后结果的决策精度最低，只有92.67%，但最后总的平均决策准确率仍较高的，有95.36%。

表4-6 十次十折多属性缺失

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 缺失属性 | | | 平均值 |
| 0，3，8 | 1，4，6 | 2，5，7 |
| min(%) | 98.53 | 97.06 | 82.35 | 92.65 |
| max(%) | 97.06 | 100 | 98.53 | 98.53 |
| avg(%) | 96.19 | 97.22 | 92.67 | 95.36 |

本节实验，主要通过本文方法在Breast Cancer数据集上进行测试，首先测试了Breast Cancer数据集中含有的16条缺失数据，其中15条数据能够正确分类，只有一条数据分类错误；然后采用十次十折交叉验证方法对681完整数据，取90%的数据进行训练DBRB，对10%的数据依次设置各个属性缺失时，检验本文方法的决策准确率，并达到较好的准确度；最后考虑多属性值缺失的情况，仍采用十次十折交叉验证，并设置测试集中有多个属性缺失，本文方法仍可以达到较好的决策准确率。

### 与DBRB方法进行比较

接下来实验将本文所提的基于模糊关联规则的方法与采用填补方法的DBRB在Iris数据集上的对比，所使用的填补方法为(1)属性均值填充方法；(2)EM填充算法；(3)KNN填补算法；(4)MICE填补算法。其中，属性均值填补方法通过统计各个属性数据并计算平均值，然后对缺失的数据进行填补；EM填充只要分为两个步骤，分别是计算完整数据的最大对数似然比，然后找到使完整数据的对数似然比最大的参数；KNN填充主要通过对完整数据集构建KD树找到相邻的K个邻居，这儿；MICE填补通过对含有不完整数据集进行多次插补，得到一个完整的数据集。使用表示本文方法。实验分别从十次两折、五折和十折交叉验证方法，折数越少用于训练的完整数据越少，取十次结果的平均表现情况作对比。

图4-4为设置单个属性缺失的实验对比结果，从图4-4(a)中可以看出，当使用二折交叉验证时，有50%的数据缺失单一属性，KNN填充方法表现最差只有75.97%的准确率，均值填充和EM方法比KNN方法来得好，但准确率不高，只有78.43%和80.03%，而MICE方法和本文方法效果最好，能使分类准确率达到90%以上接近于完整数据集的分类准确率；在图4-4(b)当使用五折交叉验证时，可以看到本文方法表示最优，能达到93.38%的准确率，而MICE方法稍逊于本文方法，准确率为92%，仍比均值、KNN和EM算法来得高；当使用十折交叉验证时，本文方法相比于二折和五折准确率来得更高，因此十折交叉验证中有90%的完整数据作为训练数据集，使得本文方法能够使用充足的数据去推断缺失数据的隶属分布，从图4-4(c)中可以看到本文方法表现最优，准确率为94.83%，比MICE方法来得更好，但MICE方法性能更加稳定。

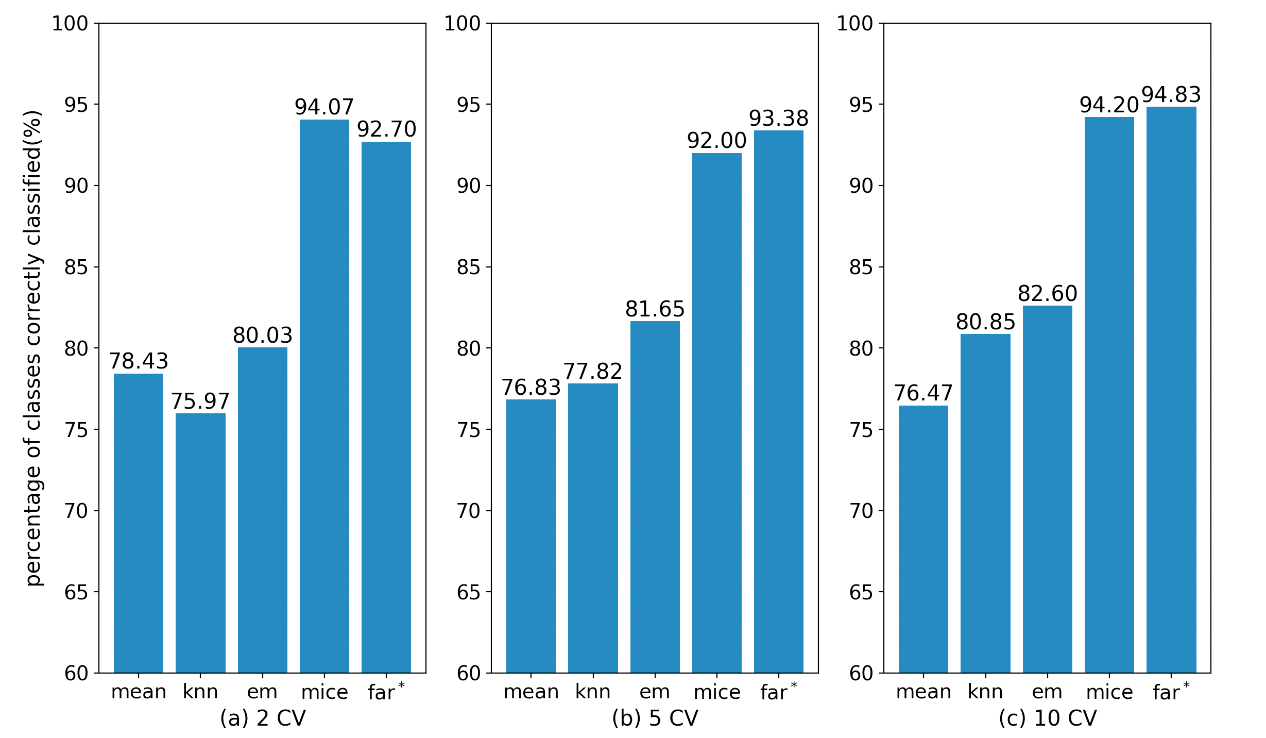


图4-4单属性缺失对比实验

图4-5为设置多个属性缺失的实验对比结果，从图4-5(a)和图4-5(b)中可以看到本文方法和MICE算法比其他方法具有更好的准确率，均能到达90%以上的分类准确率，而MICE方法表现比本文方法来得好，可以看到本文方法相比于MICE方法更容易受到完整数据的数量影响。图4-5(c)中可以看到本文方法超过了MICE方法，准确达到94.67%，同时本文方法远高于均值、KNN和EM方法。

本节实验证明了本文方法相比于使用其他算法填充的DBRB具有更好的推理准确率，在十折实验中，本文方法在单属性和多属性缺失上都比其他方法具有更好的准确率，同时本文方法比均值填充方法、EM填充方法和KNN填充方法具有更好的推理准确性。本文方法虽然在二折实验和五折的多属性缺失实验中比MICE方法中准确率来得低，但本文方法不需要再引入其他模型对缺失数据进行插补，所以本文方法能够使DBRB的系统复杂度更低。

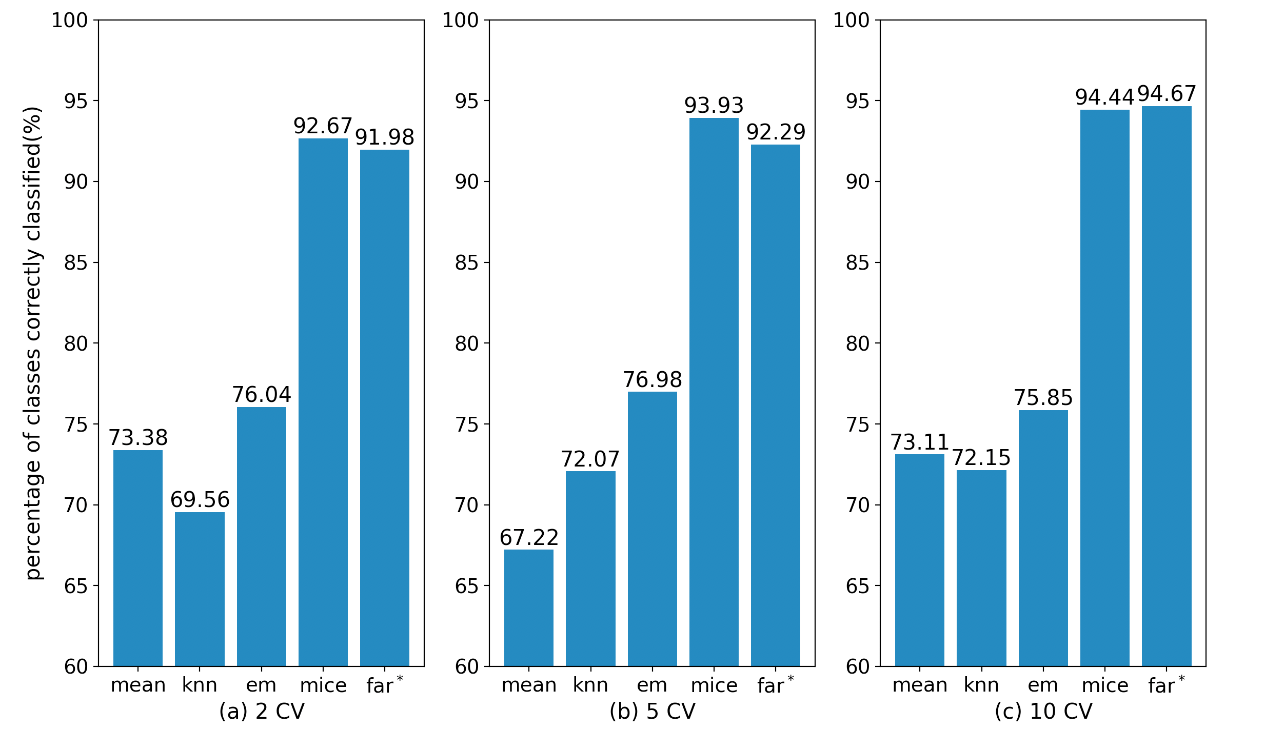


图4-5多属性缺失对比实验

### 与其他方法进行对比

本节实验将与其他方法进行对比，验证本文方法的有效性。采用十次十折交叉验证，对比实验结果的平均值。对比数据集为Breast Cancer、Iris和Transfusion数据集。对比方法选择常用的LR、CART、MLP和NBC方法。因为传统的分类方法不能处理缺失数据，因此，需要对缺失数据进行填补操作，采用本文所提的方法进行填补。

表4-7与其他方法对比实验

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 分类准确率(%) | | | 排名 |
| Breast cancer | Iris | Transfusion |
| LR | 96.41(1) | 92.38(3) | 75.87(2) | 2 |
| CART | 94.00(5) | 90.20(4) | 68.35(5) | 4.67 |
| MLP | 95.47(4) | 90.15(5) | 75.11(4) | 4.33 |
| NBC | 96.24(2) | 92.90(2) | 75.71(3) | 2.33 |
| FAR-DBRB | **96.00(3)** | **94.83(1)** | **76.01(1)** | **1.67** |

表4-7为对比实验结果，在Iris和Transfusion数据集上，基于模糊关联规则的DBRB方法比其他分类方法具有更好的分类效果；在Iris数据集上本文方法比表现第二名的方法要高1.93%；而在Transfusion数据集中，本文方法的优势不是很明显，只比第二高的LR方法要高0.14%；在Breast Cancer上本文方法效果一般，在所有方法中处于第三名，比CART和MLP方法来得较好，而LR方法表现性能最好。根据每个方法在数据集上的表现进行排名，可知本文方法在数据集中具有不错的表现，综合排名为第1。

## 本章小结

本章针对不完整信息决策问题，根据DBRB结构特点提出了基于模糊关联规则的DBRB推理方法，该方法能够有效针对输入数据中的缺失，从构建的模糊置信表示库中搜寻相关信息并进行填补。该方法能有效地针对不完整输入数据，使DBRB系统对不完整数据的推理准确得到有效地改善。为了验证本文方法，本章首先使用Breast Cancer数据集进行验证，在不同的缺失情况下，本文方法都有较好地表现；然后将本章方法与其他填补方法的DBRB方法在Iris数据集上进行对比，本文方法能够明显优于均值填充，EM填充和KNN填充的DBRB方法，与MICE方法基本相持平；最后将本章方法与传统分类方法进行对比，证明在相同的填补方法下，本文方法对不完整输入数据具有较好的表现。

# 第五章 基于DBRB的教师教学水平评估应用

## 引言

教师教学能力水平对引起学生对课程的兴趣和培养人才有着至关重要的作用[56][57][58]传统的方法是在学期末通过学生对任课教师进行打分，从而得到教师教学水平的反馈。但该方法简单且具有一定的效果，但由于学生打分过程中存在不确定主观因素将造成教师得分并不准确。同时，传统方法缺少从客观因素对教师教学水平评估。因此，本节先使用ER算法来处理学生测评中的模糊不确定性，然后通过数据库收集相关指标建立真实数据集，由于数据集中含有缺失值，因此结合第3节和第4节所提出的DBRB优化模型，对实际数据集进行评估并验证有效性。

## 基于DBRB的教师教学水平评估模型

传统的教师打分机制难以克服学生评价教师成绩过程中存在的不确定性，同时评价指标缺少客观因素过于依据学生主观评价。因此，针对上述问题，提出基于DBRB的教师教学预测模型，该模型主要分为两个部分：首先是使用ER算法融合学生的教师测评信息，得到对于教师的评价分数，然后通过分析教师相关属性构建基于DBRB的教师教学水平评估模型。

### 采用ER算法融合教师测评信息

教师测评是由学生根据学校划分的各项指标对任课教师进行评价，然后通过计算教师在各项指标的得分获得该教师的教学水平的评估结果。但该方法存在如下问题：1)教师得分依赖于学生的主观感受，学生进行评价过程中存在不确定性；2)传统的评分方式是对教师各项得分取均值作为教师最后得分，该计算方法无法很好的反应教师教学的实际水平。ER算法是在D-S证据理论的基础上发展出来，具有更广泛的应用范围，能够有效地处理信息不确定性，因此将ER算法使用于教师测评中，来反映教师更真实的教学水平。

假设教师测评中，含有项评价指标，每项指标有个评价等级，每项指标有各自的权重，且满足。如果对教师教授的某一门课程的第项评价指标的第个等级的置信度为，那么可以表示为，的计算公式如下：

|  |  |  |
| --- | --- | --- |
|  |  | 公式(5-1) |

其中，表示参与评价的学生选择评价指标等级为的人数，表示总的参与评价的学生人数。可知，，当时，代表学生全部参加教评，此时信息时完全的；当时，说明有部分学生未参加教评，信息中存在无知部分。

知晓每个评价等级的置信度后，可以通过ER算法公式(2-6)至(2-15)进行信息融合，最终得到评价结果等级的置信表示。假设评价等级效用表示为，那么可以得到最终的一个合成结果：

|  |  |  |
| --- | --- | --- |
|  |  | 公式(5-2) |

### DBRB优化方法应用

在实际问题中，数据集往往会由各种原因造成缺失和遗漏，在数据预处理中，缺失值的处理是很有必要的。在第三章和第四章针对DBRB在数据缺失的问题提出了解决方案，第三章中，针对不完整数据集问题，提出使用扩展置信规则结构的DBRB优化方法，该方法能有效的处理不完整数据集，并完整数据集上有较好的表现；在第四章中，针对不完整信息决策问题，提出使用模糊关联规则方法，该方法能有效的利用DBRB数据结构对输入数据中的缺失部分进行填补，当输入数据没有缺失项时，该方法退化为DBRB方法。可见，第三章和第四章所提的方法对完整数据的处理是缺失值处理的一个特例，因此，本节将使用第三章和第四章所提的DBRB优化方法，构建教师教学水平评估模型。

通过数据挖掘获取与教师相关的信息，例如性别、职称、任教时间、课程数、课程学时、课程选修类型、班级期末平均成绩等，其中客观因素有性别、职称、任教时间，课程属性因素有课程数、课程学时和课程选修类型，学生修习情况因素有班级期末平均成绩。通过这些因素，构建DBRB模型中的规则前件。规则后件的辨识框架为，该辨识框架由教师测评成绩的历史数据中得出，通常而言，教师测评成绩具有较高的得分，因此对其最高值和最低值所构成的区间中均匀取五个值作为评价结果效用。对于DBRB的规则库构建方法，采用第三章所提的基于扩展规则的DBRB优化方法，该方法能有效地避免参数训练所带来地系统时间开销。当使用训练好的DBRB系统进行决策时，若输入数据为不完整数据，那么采用第四章所提模糊关联规则方法对缺失项进行填补。流程如图5-1所示。

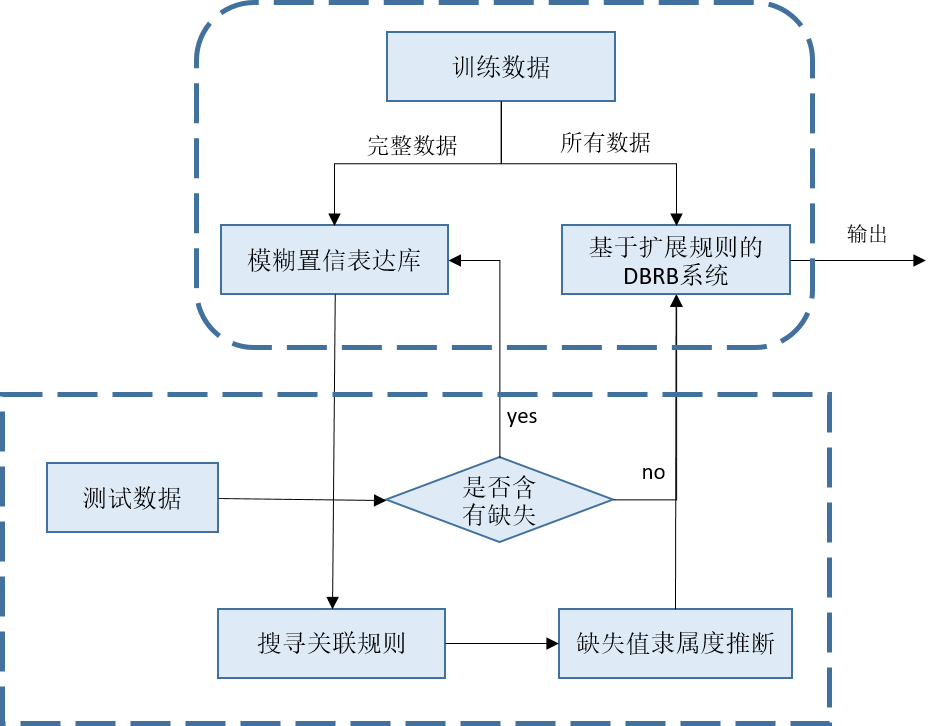


图5-1方法流程说明图

## 实例测试

本节实验利用真实数据测试教师评估模型，使用到两个数据集分别是教师测评成绩和教师教学相关数据，数据均来自福州大学2018学年第一学期相关数据。实验首先对比了ER方法与传统均值方法分别合成最终的教师得分情况，然后根据教师相关信息建立DBRB评估模型，在数据集完整和不完整的条件下与其他方法进行对比。实验环境为Intel(R) Core [i7-6700@3.40FHZ；16GB](mailto:i7-6700@3.40FHZ；16GB)内存；windows10操作系统；编程环境为Python3.8。

### 教师测评得分对比实验

本节实验采用的数据特征如表5-1所示。对数据集中选100条数据，其中每条数据含有5个评价指标，每个指标含有5个评价等级分别为不及格、及格、一般、良好和优秀，其中每个等级的效用值为40、60、70、80、100。该数据中前30条数据有同学没有参加测评，后70条数据所有的学生都进行测评。该实验分别使用均值方法和ER算法计算最后的教师测评得分，实验结果如图6-1，其中，红色圆圈表示使用的是ER算法的最后得分情况，蓝色三角表示由均值计算的得分情况。

表5-1测评数据说明

|  |  |  |
| --- | --- | --- |
| 属性个数 | 属性值 | 属性参考值 |
| 5 | 教学内容得分、教学方法得分、教学效果得分、教学素养得分、总体评价得分 | 不及格、及格、  一般、良好、优秀 |

从图中可以看到，前30个数据的评分结果差异较大，而后70个数据的评分较为一致。当存在同学没有参与测评时，采用均值方法计算时，没有将未参加测评的学生考虑在内，导致老师实际得分只能反应部分同学的评价结果，例如30个学生中只有1个学生进行测评，那么老师最终得分将会是100，显然是不合理的。ER算法的得分比均值来得更低，因为ER算法能够考虑没有进行评价学生所导致的不确定，当参加测评的学生较少时，学生在测评中的评价结果具有较低的置信度，所以教师最终得分较低，该评分方法比较符合实际情况。

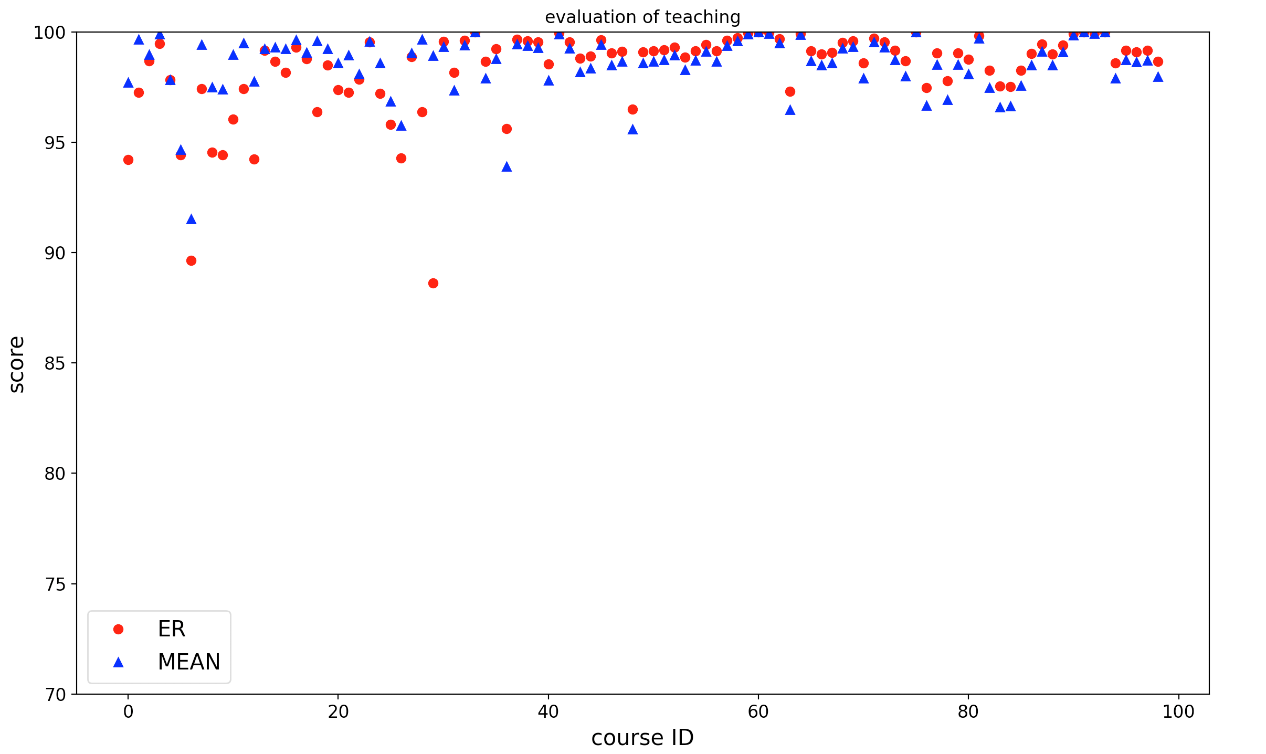


图5-2教师测评对比实验

### 基于DBRB的评估方法与其他方法的比较

表5-2为所收集的教师在学期授课中的相关信息，其中，该数据集含有部分缺失数据，缺失数据主要集中在教师开始任教时间、授课课程学时以及授课课程选修类型。为了更好地进行对比实验，本节实验首先将数据集中缺失的数据不考虑在内，使用本文方法和其他方法在完整数据集上进行对比，然后再对缺失的数据集进行对比。

表5-2评估数据说明

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 数据个数 | 属性个数 | 含缺失项属性(个) | | 输出值 |
| 2514 | 7 | | 属性2(945)、属性5(96)、属性6(8) | 数值0~100 |

实验对比方法为支持向量回归方法(svr)、决策树回归方法(dtc)和多层感知器方法(mlp)。由于传统的方法无法直接处理缺失的数据，因此实验开始前先对缺失数据进行填补，在4.5.2节中可知，MICE方法相比其他填补数据方法具有更好效果，因此本节实验采用MICE方法对数据集中缺失的部分进行填补。为了方便，本文方法简称为e-dbrb。检验指标为平均绝对误差(MAE)和均方根误差(RMSE)。实验采用十折交叉验证。

#### 完整数据集下的对比实验

图5-3为十折交叉验证中每次交叉验证的结果，可知本文方法在MAE中具有较低的表现，与mlp方法基本相持平，比svr和dtc方法具有更小的平均绝对误差。从图5-4可知，本文方法与dtc和mlp方法具有相近的RMSE值。可知本文方法相比于传统的回归模型具有相近的表现性能，在完整数据集中，本文方法能够有效的对教师教学水平进行评估。

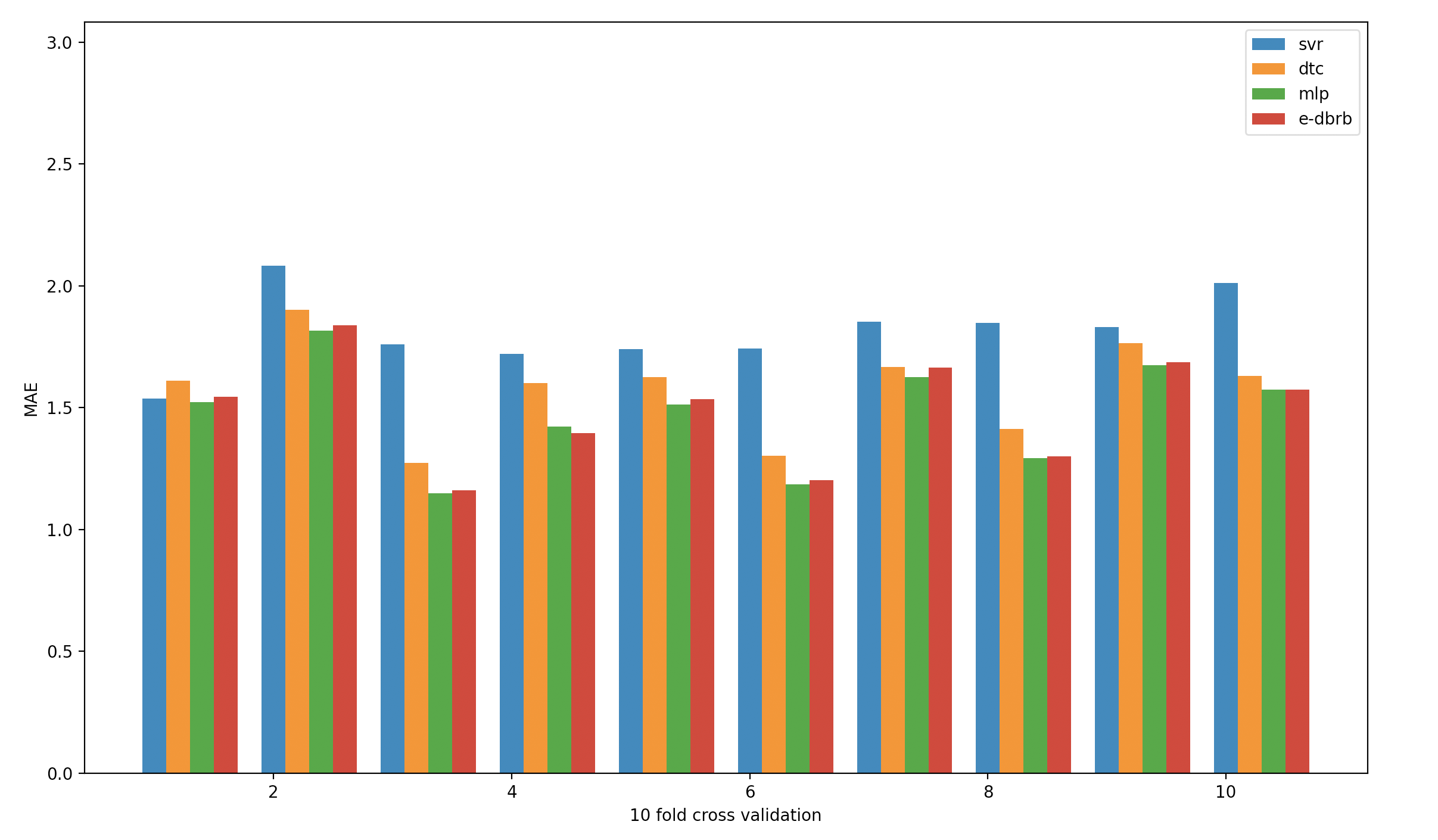


图5-3 在完整数据集的MAE对比实验

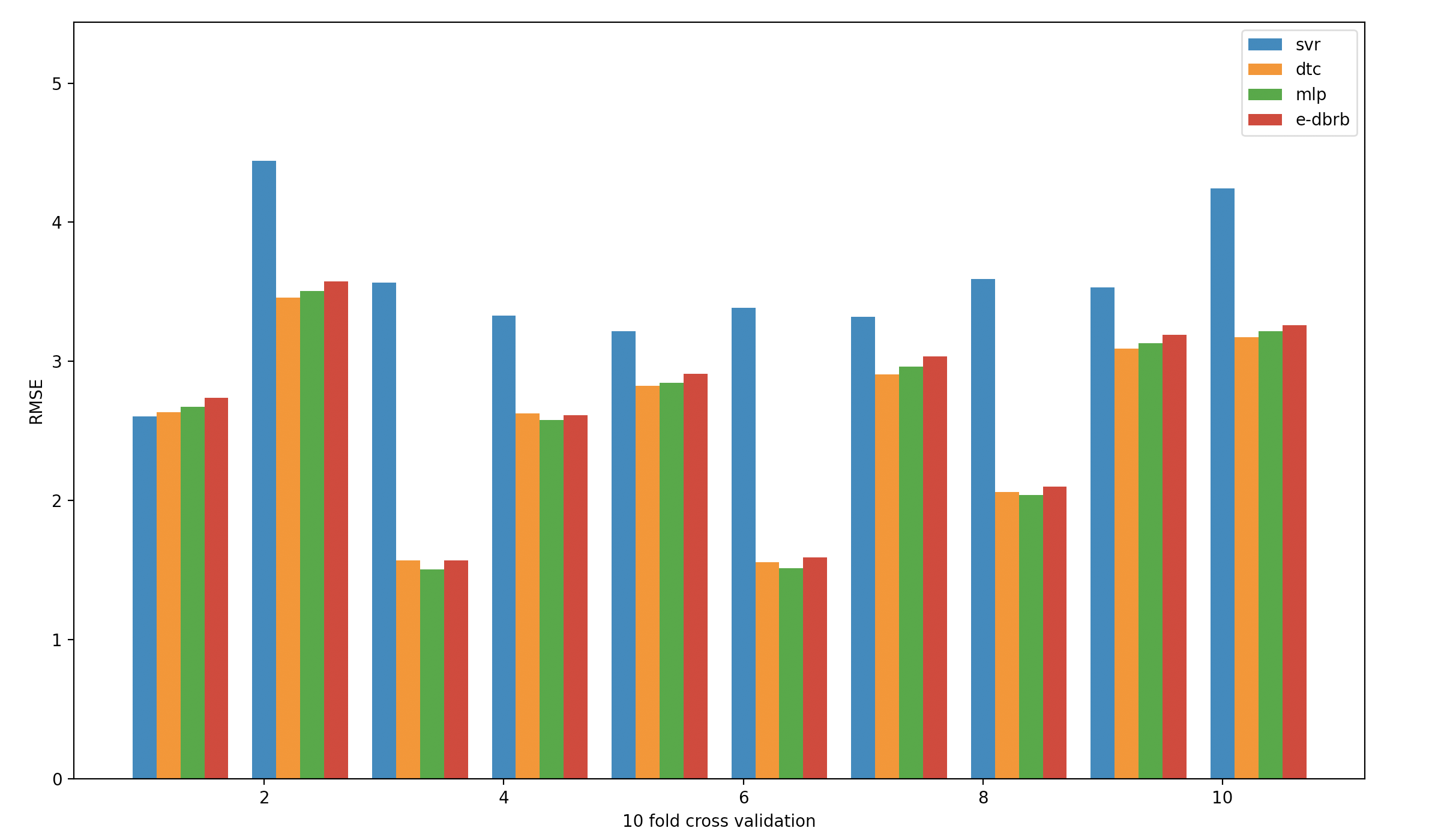


图5-4 在完整数据集的RMSE对比实验

表5-3为十折交叉验证结果取均值的对比结果，从表中知，本文方法在MAE中具有较小的误差，与表现最好的mlp相差0.018左右，比svr和dtc低5.8%和14.9%。在RMSE指标中，本文方法表现排名第三，比mlp和dtc表现来得差，但RMSE值十分接近，而svr表现来得最差，表明svr在十折交叉验证中的波动性比较大，本文方法与mlp和dtc具有相近的评估结果。

表5-3完整数据对比实验

|  |  |  |
| --- | --- | --- |
| 方法 | MAE | RMSE |
| SVR | 1.750020 | 3.509095 |
| DTC | 1.552010 | 2.609862 |
| MLP | 1.488853 | 2.605198 |
| **E-DBRB** | **1.490142** | **2.664352** |

#### 不完整数据集下的对比实验

图5-5和图5-6为十折交叉验中每一次交叉验证的对比结果。从图5-5可知，本文方法在十折交叉验证中，每次的表现均比其他方法来得好，具有更低的MAE，说明本文方法的表现性能与真实值存在较小的差距，具有更好的准确率。而图5-6的RMSE指标上，本文方法表现一般，比svr方法表现得好，但比dtc和mlp方法表现得差，在十折交叉验证中，平均表现性能处于第三名。

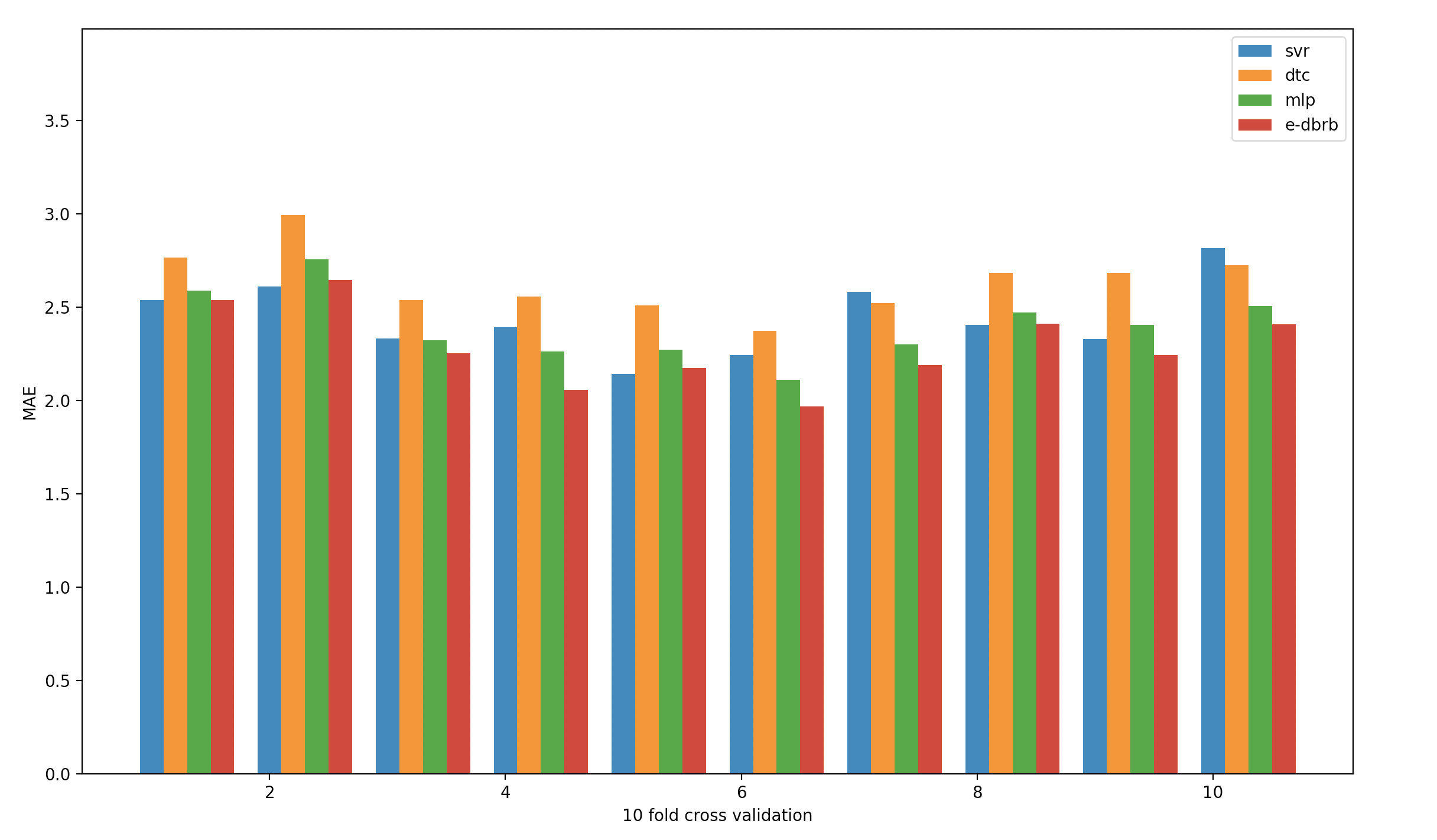


图5-5 在不完整数据集的MAE对比实验

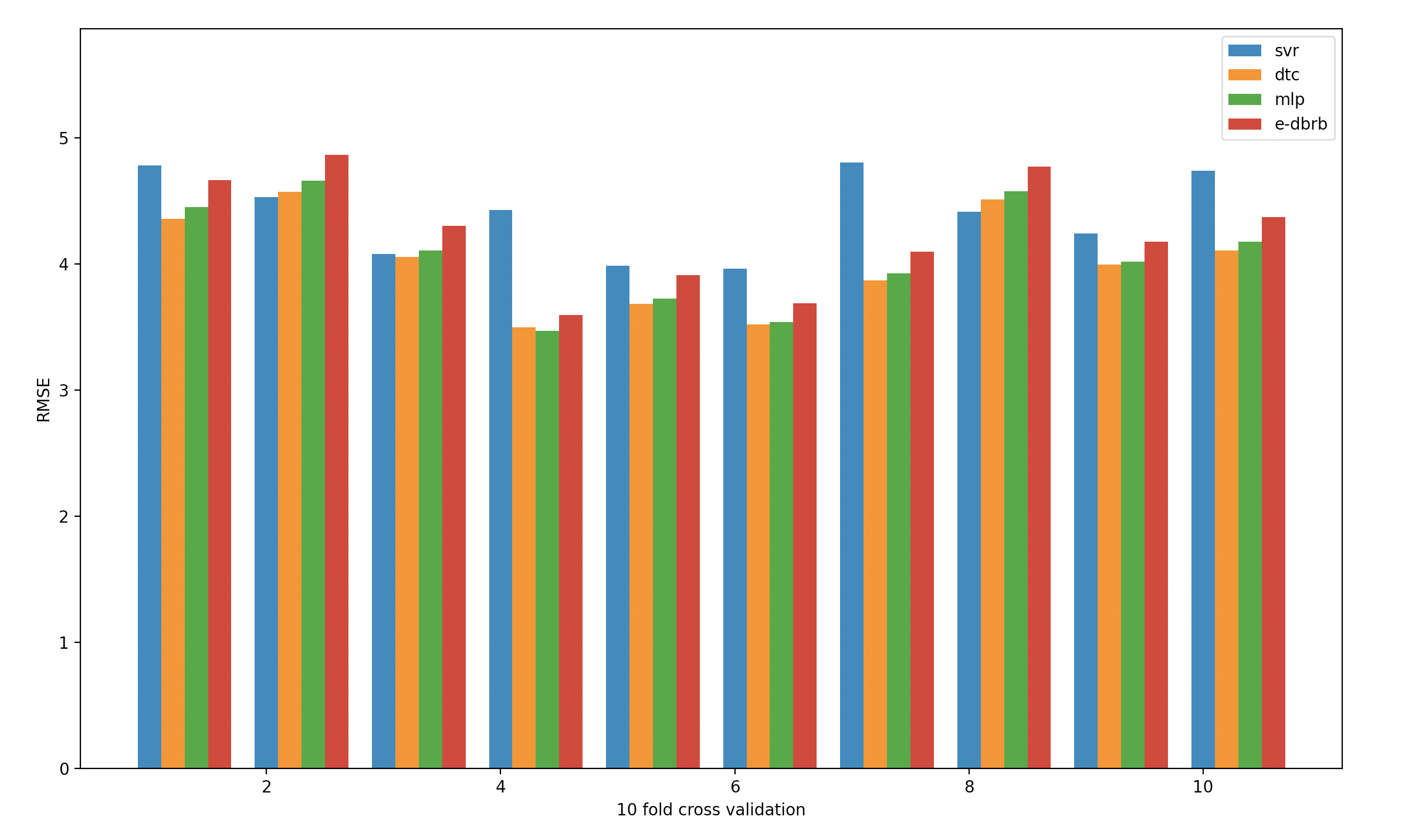


图5-6 在不完整数据集的RMSE对比实验

表5-4为十折交叉验证取均值的对比结果，从表中可知，本文方法的MAE值比其他方法来得小，比第二名方法SVR低4.7%，表明了本文方法的估计值与真实值之间存在的误差更小。而在RMSE中，本文方法表现比DTC好，但比其余两种方法表现来得差，这说明了本文方法的评估结果波动性较大，容易出现异常值，方法欠缺稳定的表现能力。相比完整数据集与不完整数据集上的实验，在MEA和RMSE两个指标中均出现更大的误差和较低的拟合效果。

表5-4不完整数据对比实验

|  |  |  |
| --- | --- | --- |
| 方法 | MAE | RMSE |
| SVR | 2.400613 | 4.443243 |
| DTC | 2.637964 | 4.024410 |
| MLP | 2.428186 | 4.247090 |
| **E-DBRB** | **2.286976** | **4.267413** |

本节实验对本文方法在教师教学水平评估问题上进行验证，通过真实数据集进行测试，并与其他回归模型进行对比。从实验结果可知，本文方法在完整数据集上具有良好的表现性能，具有较低的误差和较好的拟合效果，而在不完整数据集上，本文方法在MAE指标中表现最好，说明本文方法的评估值更为准确，但在RMSE指标中表现一般，说明本文方法仍欠缺较好的鲁棒性，评估结果波动较大。缺失数据集难免会出现信息损失，实验的四个模型在不完整数据集上的误差均比完整数据集上的表现来得差，但本文方法能够在保持完整数据集上有较好的表现性能的同时，在不完整数据集上比其他方法具有更小的误差。

## 本章小结

本章对前两章提出的DBRB模型进行应用，在教师教学水平评估领域中建立DBRB模型对教师教学水平进行评估，该模型融合了前两种方法处理缺失数据的特点，使得该模型可以同时处理不完整数据集合不完整信息决策问题。首先该使用ER算法处理带有不确定信息的教学测评情况，获取接近教师真实水平的教学测评得分；然后使用提出的教学水平评估模型对完整和不完整的数据集进行验证。在真实的教师测评数据集中，当数据集完整时，本文方法表现一般，和其他方法相比，性能位居第二；而当数据集不完整时，本文方法的评估值比其他方法的评估结果更加接近真实值，具有更好的性能。本文方法除了在完整数据集上有良好的表现，还能处理含有缺失值的数据集，因此该方法更具有实际应用意义。

# 总结与展望

Yang等[35]最先提出了置信规则库推理方法，在传统的IF-THEN规则结构上引入了置信分布框架，并以该规则结构为基础提出了基于证据推理的RIMER方法，该方法能处理各种不完整性、不确定性、模糊性信息，并成功应用于输油管道检漏和桥梁风险评估等实际生活生产领域中。基于析取范式的置信规则库由Yang提出并进行讨论，而后由Chang等[19]对其激活权重进行改进和提出优化参数方法，并成功运用于分类问题。早先DBRB模型依赖于专家知识进行参数设置，随着不断的研究深入，DBRB系统能够通过历史数据进行参数学习，使DBRB系统能够自动化的完成参数初始化。然而DBRB系统依赖于历史数据的质量，含有缺失值的数据将会导致DBRB系统性能的下降，因此对于数据缺失的问题是DBRB中需要解决一个重要问题。所以，本文针对缺失数据问题分别提出基于扩展规则的DBRB优化方法和基于模糊关联规则的DBRB优化方法，并在教师教学水平评估的实际问题上进行应用，具体研究内容包括如下三点：

(1) 在DBRB中不完整数据集会导致参数学习出现偏差，同时，传统的CBRB系统无法处理含有缺失项的数据集，而DBRB对缺失的数据具有一定的兼容性，但无法保证准确率，因此，本文针对上述问题，提出基于扩展规则的DBRB优化方法，该方法借鉴了扩展置信规则库的构建方法，通过不完整数据集生成不完整扩展置信规则，并引入衰减因子对规则权重进行调节，使规则权重更合理地参与信息融合步骤中。该方法避免了传统DBRB中繁复的参数学习过程，同时能有效地利用不完整数据中的信息，使在不增加系统复杂度的前提下有效提高不完整数据集下DBRB系统的推理准确率。为了验证本文方法的有效性，首先在完整数据集上将本文方法与其他方法进行对比，然后在不同的缺失模式和缺失率下进行对比，本文方法均具有较好的表现性能。

(2) 不完整信息决策问题是数据缺失的另一种类型，传统的方法无法有效的利用已有的信息，针对这种情况，本文提出了基于模糊关联规则的DBRB优化方法，该方法首先使用高斯隶属函数优化个体匹配度的计算，然后利用训练数据集构建模糊置信表示库，当决策的数据是不完全时，可以通过模糊置信表示库搜寻关联规则，并对缺失数据进行推断。为了说明本文方法的有效性，首先将本文方法在公共数据集上进行验证有效性，然后将本文方法与传统的填补方法进行对比，本文方法与性能最好的MICE方法具有相近表现，最后将本文方法与其他分类方法进行对比，均具有不错的表现。

(3) 使用本文所提的优化方法针对教师教学水平评估问题进行研究应用。教师教学水平评估所使用的数据集来自于福州大学2018学年01学期数据，本文首先针对传统的打分机制的缺点，选择使用ER算法获取学生对教师的测评成绩，因为实际数据集中含有部分缺失值，所以使用本文的DBRB优化方法进行建模，最后与其他回归模型进行比对，证明本文方法的有效性。

本课题关于DBRB的优化及应用研究仍有不足需要改善，未来的研究工作展望如下：

(1) 目前，基于扩展规则的DBRB优化方法中，需要针对数据集的特点确定前提属性参考值的个数，该步骤尚未实现自动化设置，在实际应用中进行预实验搜寻最优参考值个数值，这将对系统的性能造成负担。因此，未来的研究可以考虑实现参数自动化设置，通过参数学习或其他方法完成对该参数的设置，使该方法更加完善。

(2) 基于隶属度的关联规则推断方法虽然能够有效的处理缺失数值，但需要遍历整个模糊置信表示库，这将造成系统性能上的低效。因此，接下来的研究将考虑如何更高效地从缺失数据中构建关联规则，抑或是提出更好地方法解决该问题。

(3) 本文的DBRB优化方法在教师教学水平评估中进行应用，但仍有许多可以进行改进的地方，例如对数据集属性进行属性约简，判断属性与结果的相关程度，以及挖掘更多相关属性的数据，使在教师教学水平评估中的应用更有效。

# 参考文献

1. 周志杰. 置信规则库专家系统与复杂系统建模[M]. 科学出版社, 2011.
2. Mohan S , Arumugam N . Expert system applications in irrigation management: an over view[J]. Computers and Electronics in Agriculture, 1997, 17(3):263-280.
3. Chomsky N . Rules and representations[J]. 1980.
4. Sun R . Robust reasoning: integrating rule-based and similarity-based reasoning[J]. Artificial Intelligence, 1995, 75(2):241-295.
5. Kyburg H E . Probabilistic Reasoning in intelligent systems[J]. Journal of Philosophy, 1988, 88(8):434-437.
6. Zadeh L A . Fuzzy logic[J]. Computer, 1988, 21( 4), 83–93.
7. Spiegelhalter D J . Bayesian analysis in expert systems[J]. Statistical science, 1993, 8(3):219-247.
8. Chen W T, Saif M. A novel fuzzy system with dynamic rule base[J]. IEEE Transactions on Fuzzy Systems, 2005 13(5), 569-582.
9. Yang J B , Liu J , Wang J , et al. Belief rule-based inference methodology using the evidential reasoning approach-RIMER[J]. IEEE Transactions on Systems Man & Cybernetic Part A Systems & Humans, 2006, 36(2):266-285.
10. Zadeh L Z. Fuzzy set[J]. Information Control, 1965, 8(3):338-353.
11. Hwang, C.L, K. Yoon, Yoon, K. Multiple Attribute decision making: methods and applications[M]// Fuzzy Multiple Objective Decision Making :.
12. 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.
13. 杨隆浩, 蔡芷铃, 黄志鑫, 等. 出租车乘车概率预测的置信规则库推理方法[J]. 计算机科学与探索, 2015, 9(8):985-994.
14. Wang Y M , Elhag T M S . A comparison of neural network, evidential reasoning and multiple regression analysis in modelling bridge risks[J]. Expert Systems with Applications, 2007, 32(2):336-348.
15. Jiang J, Li X, Zhou Z, et al. Weapon system capability assessment under uncertainty based on the evidential reasoning approach[J]. Expert Systems with Applications, 2011, 38(11):13773-13784.
16. Yang J B , Wang Y M , Xu D L , et al. Belief rule-based methodology for mapping consumer preferences and setting product targets[J]. Expert Systems with Applications, 2012, 39(5):4749-4759.
17. Kong F , 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.
18. 方志坚, 杨隆浩, 傅仰耿, 等. 基于置信规则库推理的多属性双边匹配决策方法[J]. 南京大学学报(自然科学), 2016, 52(4): 672-681.
19. Chang L , Zhou Z J , You Y , et al. Belief rule based expert system for classification problems with new rule activation and weight calculation procedures[J]. Information Sciences, 2016, 336(C):75-91.
20. Chang L , Ma X , Wang L , et al. Comparative analysis on the conjunctive and disjunctive assumptions for the belief rule base[C]// International Conference on Cyber-enabled Distributed Computing & Knowledge Discovery. IEEE Computer Society, 2016.
21. Yang L H , Wang Y M , Chang L L , et al. A disjunctive belief rule-based expert system for bridge risk assessment with dynamic parameter optimization model[J]. Computers & Industrial Engineering, 2017, 113:459-474.
22. Chang L L , Jiang J , Sun J B , et al. Disjunctive belief rule base spreading for threat level assessment with heterogeneous, insufficient, and missing information. Information Sciences, 2019, 476:106-131.
23. Zhu W , Hou P Z , Chang L L , et al. Disjunctive belief rule base optimization by ant colony optimization for railway transportation safety assessment[C]// 2019 Chinese Control And Decision Conference (CCDC). 2019.
24. Yang J B , Liu J , Xu D L , et al. Optimization models for training belief rule-based systems[J]. IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans, 2007, 37(4):569-585.
25. 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.
26. 常瑞, 张速. 基于优化步长和梯度法的置信规则库参数学习方法[J]. 华北水利水电大学学报, 2011, 32(1): 154-157.
27. 王韩杰, 杨隆浩, 傅仰耿, 等. 专家干预下置信规则库参数训练的差分进化算法[J]. 计算机科学, 2015, 42(5): 88-93.
28. 苏群, 杨隆浩, 傅仰耿, 等. 基于变速粒子群优化的置信规则库参数训练方法[J]. 计算机应用, 2014, 34(8): 2161-2165.
29. Chang L , Zhou Y , Jiang J , et al. Structure learning for belief rule base expert system: A comparative study[J]. Knowledge-Based Systems, 2013, 39(feb.):159-172.
30. 杨隆浩, 王晓东, 傅仰耿. 基于关联系数标准差融合的置信规则库规则约简方法[J]. 信息与控制, 2015, 44(1): 21-28, 37.
31. 王应明, 杨隆浩, 常雷雷, 等. 置信规则库规则约简的粗糙集方法[J]. 控制与决策, 2014, 29(11): 1943-1950.
32. Zhou Z J , Hu C H , Yang J B , et al. A sequential learning algorithm for online constructing belief-rule-based systems[J]. Expert Systems with Applications, 2010, 37(2):1790-1799.
33. Nozaki K , Ishibuchi H , Tanaka H . Adaptive fuzzy rule-based classification systems[J]. Fuzzy Systems IEEE Transactions on, 1996, 4(3):238-250.
34. Wang Y M , Yang L H , Fu Y G , et al. Dynamic rule adjustment approach for optimizing belief rule-base expert system[J]. Knowledge-Based Systems, 2016, 96(Mar.15):40-60.
35. 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.
36. Yang L H , Wang Y M , Lan Y X , et al. A data envelopment analysis (DEA)-based method for rule reduction in extended belief-rule-based systems[J]. Knowledge-Based Systems, 2017, 123: 174-187.
37. Alberto C, Liu J, Wang H, Anil K. A new dynamic rule activation method for extended belief rule-based systems [J]. IEEE Transactions on knowledge and data engineering, 2015, 7(4):880-888.
38. Lin Y Q, Fu Y G, Su Q, et al. 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.
39. 叶青青, 杨隆浩, 傅仰耿,等. 基于改进置信规则库推理的分类方法[J]. 计算机科学与探索, 2016, 10(5):709-721.
40. Chang L , Chen Y , Hao Z , et al. Indirect disjunctive belief rule base modeling using limited conjunctive rules: Two possible means[J]. International Journal of Approximate Reasoning, 2019, 108(may):1-20.
41. Wang X , Sun J , You Y , et al. A new approach for disjunctive belief rule base construction with[C]// 2019 Chinese Control and Decision Conference (CCDC). 2019.
42. Chang L L , Zhou Z J , Chen Y W , et al. Belief rule base structure and parameter joint optimization under disjunctive assumption for nonlinear complex system modeling[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2018, 9:1-13.
43. Wang Y M , Yang J B , Xu D L , et al. Environmental impact assessment using the evidential reasoning approach[J]. European Journal of Operational Research. 2006, 174(3):1885-1913.
44. 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.
45. 王小燕,孙建彬,赵青松,常雷雷,邹志刚.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.
46. Lin Y Q , Fu Y G , Su Q , et al. 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.
47. Yang, L H , Wang Y M, Fu Y G . A consistency analysis-based rule activation method for extended belief-rule-based systems[J]. Information Sciences An International Journal, 2018.
48. Yang L H , Liu J , Wang Y M , et al. New activation weight calculation and parameter optimization for extended belief rule-based system based on sensitivity analysis[J]. Knowledge and Information Systems, 2018:1-42.
49. Razavi-Far R , Cheng B , Saif M , et al. Similarity-learning information-fusion schemes for missing data imputation[J]. Knowledge-Based Systems, 2020, 187(Jan.):104805.1-104805.13.
50. 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(04):51-59.
51. Li S H , Feng J Y , He W. Health assessment for a sensor network with data loss based on belief rule base[J]. IEEE Access. 2020, 8, 126347-126357.
52. Zadeh L A . Fuzzy sets[J]. Information & Control, 1965, 8(3):338-353.
53. Zhang J, Fu Y G, Chen N N , Wu Y J . Activation method for disjunctive belief rule base using principle of inclusion and exclusion[C]// 2018 international Conference on Cloud Computer, Big Data and Blockchain.2018.
54. Storn R . Differential evolution-a simple and efficient heuristic for global optimization over continuous space[J]. Journal of Global Optimization, 1997, 11.
55. Mangasarian O L , Wolberg W H . Cancer diagnosis via linear programming[C]// Siam News. 1990.
56. 李丽菊. 教师测评知识:有效教学的重要保障[J]. 现代教育科学, 2013.
57. 肖杨田, 潘鸣威. 教师测评素养国际研究热点与趋势分析(2005-2019)--基于SSCI期刊论文的可视化分析[J/OL]. 外语研究, 2020(05):43-51
58. Donaldson M . Multidisciplinary perspectives on teacher evaluation: Understanding the research and theory[M].Taylor and Francis:2020-08-27.

# 致谢

时光如梭，转眼研究生的两年半时光已经快要结束，回首这段难忘的时光，经历了许多挫折与迷茫，也有些许收获和成长。课堂中，在老师的教授下，学习了更多的知识；实验室中，与同学的和睦相处，获得了珍贵的友谊；BRB研究小组中，在导师的教导下，取得了宝贵的科研成果。面对即将结束的研究生生涯，心中充满了不舍之情，真诚地向那些给予我帮助、鼓舞、支持和指导的老师、同学、家人和朋友表达自己最真挚的感谢。

首先，我要感谢我的导师傅仰耿副教授。感谢导师对我学术上的指导，在科研的海洋中为迷茫的我指明方向。从科研入门到小组汇报，再从论文选题、到论文写作，导师始终耐心且细心地进行指导，让我了解自己需要改正的问题，同时导师提供了众多讲座论文资料让我能够从中获得额外的知识收获。导师多次教导我们要坚持学术诚信，坚守学术道德底线，导师严谨的学术作风使我受益匪浅，对我的科研工作具有重要的影响。

然后，还要感谢实验室的吴英杰教授、王一蕾副教授和孙岚老师，感谢他们为实验室无私的付出，才能让我在一个良好的环境下进行科研与学习。感谢吴英杰教授让我参与到实验室项目中，让我能够磨练自己，学习新的技能并获得成长。感谢王一蕾副教授对我的教诲，让我知道自己的不足与缺点。感谢孙岚老师对我在实验室中担任的工作给予了鼓舞与肯定，让我能够更加自信。

同时，感谢BRB研究小组的每位成员：刘莞玲学姐、张婕学姐、苏曼娜学姐、黄宏运学长、陈楠楠学长、方炜杰学长、庄金惠、尹泽锋、管宇、郑铭鸿、叶己峰、兰霖、陈龙江等同学。感谢他们每周在组会上进行汇报与分享，让我有机会了解更多的科研最新进展。感谢郑铭鸿和陈龙江同学帮我检查论文，提出了一些宝贵的修改意见。

其次，感谢实验室的小伙伴在我的学习科研上给予了关心与支持，感谢邱瑜冰同学与程祺同学在实验室中不断激励我前行，在我遇到困难时给予了帮助。感谢宿舍室友在生活上给予了我包容与理解，感谢朱伟大同学在研究生生活中的陪伴，给了我许多有益的建议。

最后感谢家人的鼓励和陪伴，在我的读研生活中给予了重要的支持，让我能够顺利完成学业。

# 个人简历

|  |  |
| --- | --- |
| **姓名：** | 刘永裕 |
| **出生年月：** | 1995年10月 |
| **性别：** | 男 |
| **籍贯：** | 福建泉州 |

|  |  |  |
| --- | --- | --- |
| **学习及工作经历：** | | |
| 2018.09-2021.02 | 福州大学数学与计算机科学学院 | 工程硕士研究生 |
| 2014.09-2018.07 | 郑州大学数学与统计学院 | 理学学士学位 |

# 在学期间的研究成果及发表的学术论文

在读期间已发表和录用的论文：

第一作者(1篇)：

1. 刘永裕,巩晓婷,方炜杰,傅仰耿. 数据缺失的扩展置信规则库推理方法[J]. 计算机研究与发展.2021,58.(已录用)

参与的科研项目及成果：

1. 国家自然科学基金项目(61773123)
2. 福建省自然科学基金项目(2019J01647)
3. 工信部工业互联网创新发展工程项目(TC19083WB)