引用格式： 朱勋程，赵忆宁，李宁，等.D-S证据理论驱动的统一社会信用代码数据融合补全方法[J].标准科学，2025(9):52-58.
ZHU Xuncheng,ZHAO Yining,LI Ning,et al. Fusion and Completion Method for Data of Unified Social Credit Code Driven by D-S Evidence Theory [J].Standard Science,2025(9):52-58.

D-S 证据理论驱动的统一社会信用代码数据融合补全方法

朱勋程　赵忆宁*　李 宁　黄典一　李瑜敏　冯斯佑　杨 俊　杨 敏
（云南省标准化研究院）

摘　要：【目的】立足于信用体系的完整性与有效性需求，开展法人和其他组织统一社会信用代码（以下简称“统一代码”）缺失数据补全方法研究，以提升信用体系的整体效能与可信度。【方法】采用均值、回归分析、K-最近邻、随机森林、NGboost等多种方法对缺失数据进行补全；随机删除完整数据的部分属性数据，将不同算法的数据补全精度作为其可信度；通过证据理论根据不同算法的可信度和其补全数据值融合为最终补全数据值。【结果】所提出的方法能够综合不同补全算法的优点，并能够对最终补全的数据可信度进行量化。【结论】该方法有效提升了统一社会信用代码数据的完整性和准确性，进一步增强了信用体系的有效性和可靠性。

关键词： 统一社会信用代码；数据补全；D-S证据理论
DOI编码： 10.3969/j.issn.1674-5698.2025.09.007

Fusion and Completion Method for Data of Unified Social Credit Code Driven by D-S Evidence Theory

ZHU Xuncheng　ZHAO Yining*　LI Ning　HUANG Dianyi　LI Yumin
FENG Siyou　YANG Jun　YANG Min
(Yunnan Institute of Standardization)

Abstract: [Objective] The study aims to improve the overall efficiency and credibility of the credit system, and proposes a method of completing the missing data of the unified social credit code. [Methods] Mean calculation, regression analysis, K-nearest neighbors, random forest, NGboost etc. are used to complete the missing data. The study randomly deletes part

基金项目： 本文受云南省市场监督管理局科技计划项目“基于机器学习的统一社会信用代码数据补全关键技术研究”（项目编号：2023YSJK01）；云南省科技厅科技计划项目“技术创新人才培养对象项目朱勋程”（项目编号：202405AD350104）资助。

作者简介： 朱勋程，本科，正高级工程师，研究方向为大数据分析和标准数字化转型。
赵忆宁，通信作者，本科，高级工程师，研究方向为标准信息检索、统计、分析及应用。
李宁，硕士，高级工程师，研究方向为标准化研究与应用。
黄典一，硕士，工程师，研究方向为标准化研究和标准数字化管理。
李瑜敏，硕士，助理工程师，研究方向为标准化研究和数据分析。
冯斯佑，硕士，助理工程师，研究方向为标准数字化。
杨俊，硕士，高级工程师，研究方向为标准化和信息化研究及管理。
杨敏，硕士，工程师，研究方向为数字标准技术研究。

of the attribute data from the complete data, using the data completion accuracy of different algorithms as their credibility. The credibility of different algorithms and their completion data values are fused to form the final completion data values through the evidence theory. [Results] The proposed method can combine the advantages of different completion algorithms and quantify the credibility of the final completion data. [Conclusion] The proposed method can improve the completeness and accuracy of the unified social credit code data effectively, and further enhance the effectiveness and reliability of the credit system.

Keywords: unified social credit code; data completion; D-S Evidence Theory

0 引言

2015年6月11日，国务院发布《法人和其他组织统一社会信用代码制度建设总体方案》（国发〔2015〕33号）[1]，标志着我国统一代码制度全面实施。统一代码的唯一性是指18位统一代码及其9位主体标识码（组织机构代码）在全国范围内是唯一的，即1个主体只能拥有1个统一代码[2]。基于这种唯一性，统一代码能够进行身份核验，开展代码主体验证服务，识别信息中的代码主体，跟踪代码主体动态信息，开展代码主体跟踪监测服务[3]。此外，统一代码可以将每个经营主体分散在各地区、各部门、各领域的信用记录归集整合到自己的名下，实现跨地区、跨部门、跨领域的信息共享，可以说是“一次采集，多方利用”，形成完整统一代码档案[4]。在失信联合惩戒中，统一代码数据发挥着核心作用，它不仅提高信用监管的效率和准确性，还为构建公平诚信的市场环境提供有力支持[5]。

随着现代市场经营环境趋于复杂，统一代码同信用监管的深度绑定对统一代码数据质量提出更高要求[6]。然而，在统一代码制度的实施与数据管理过程中，部分数据由于历史久远且多为人工录入，存在数据缺失问题，这使得整体数据质量受到影响。若不对这些缺失数据进行补全，将难以有效开展数据挖掘等分析任务，进而影响统一代码数据的广泛应用。

因此，本文提出了一种基于D-S证据理论的统一代码缺失数据多策略融合补全方法，旨在有效整合多种补全算法的优势。通过综合应用这些算法进行补全值的融合，并对最终结果进行可信度评估，提升统一代码数据补全的准确性和可靠性。

1 相关工作

1.1 数据补全技术

数据补全技术在机器学习应用场景中引起广泛关注。数据补全技术旨在填补数据集中缺失的值，提高数据的完整性和分析的准确性，这对于确保模型的有效性和决策的可靠性至关重要。目前，存在多种传统统计和机器学习技术用于处理缺失值，包括均值插补、回归插补、K-近邻算法（K-Nearest Neighbors Algorithm，KNN）及基于集成学习的方法等[7]。Hussain等[8]引入了一种新颖的特征工程框架——NGBoost，用于检测电力消耗数据中的欺诈行为。该框架通过结合概率增强和梯度提升技术，能够有效识别和建模电力消耗数据中的异常模式，从而提高欺诈检测的准确性和灵活性。Karamti等[9]针对预测宫颈癌数据集中缺失值的挑战，提出了一种自动化系统，利用KNN的合成少数类过采样技术处理缺失值，在数据缺失的情况下仍能实现高准确率的宫颈癌预测。Lee等[10]评估了基于多重随机森林算法的填补方法在处理协变量中缺失数据方面的应用。Khan等[11]对流行的链式方程多元插补算法进行了扩展，提出了2种变体以填补分类和数值数据。Bai等[12]通过子空间回归评估样本间的相关性，预测潜在的缺失值，然后在自编码器框架内利用这些预测进行插值。Liu等[13]用前向数据补全方法处理单次和多次访问中的缺失和异常数据。Lena等[14]分析了在不同缺失模式下，针对甲基化数据的不同补全方法的表现。

Razavi-Far等[15]提出了数据补全的相似性方案，以提高缺失值填补的准确性和鲁棒性。该研究通过引入相似性学习机制，能够有效捕捉数据之间的潜在关系，从而在缺失数据的上下文中提供更全面的信息。这一方法不仅增强了补全结果的可靠性，还为处理复杂数据集中的缺失值提供了新的思路和解决方案。

上述研究均将数据补全技术与各自的研究领域紧密结合。然而，以上研究通常依赖于单一的补全方法，缺乏对多种补全策略的综合运用。本文将针对统一代码数据缺失问题提出有效解决方案，旨在进一步提升数据补全的准确性。

1.2 统一代码数据分析

本文进行研究所用数据来源于云南省统一代码数据库。该数据库涵盖了机构的登记信息，包括企业名称、统一社会信用代码、行业类型、注册资金、从业人数、注册地址等多种类型的数据，部分参数类型如表1所示。

表1 企业部分数据类型示例

参数名称	参数类型	字段名称
qymc	String	企业名称
uniscid	String	统一社会信用代码
ywlx	String	登记业务类型
jyfw	String	经营范围
jyzt	int	经营状态
business_type	String	行业类型
Zczj	int	注册资金
Cyrs	int	从业人数
Zcdz	String	注册地址

本文进行实验研究所涉及企业数据的示例如表2所示，由于原始数据存在部分字段缺失和数值不准确等问题，需要通过数据补全技术提升数据完整性和可靠性。

表2 企业数据示例

企业名称	注册地址	从业人数	注册资金
昆明×××有限公司	云南省昆明市×××号	30	5
昭通×××有限责任公司	云南省昭通市×××镇	21	5000
云南×××有限公司	云南省昆明市×××区	121	40
云南省×××协会	昆明市×××号	30	5
……	……	……	……

2 统一代码缺失数据多策略融合补全方法

2.1 整体框架

传统统一代码数据补全方法多采用插值法、平均值/中值法、最近邻法、回归分析和基于规则等单一策略。然而，这些方法存在局限性：它们依赖单一技术，缺乏对多种策略的综合运用和可信度评估，且在处理不同缺失模式时难以保证数据质量。为解决这些问题，本文提出了一种基于D-S证据理论的统一代码缺失数据多策略融合补全方法，其方法框架如图1所示。该方法采集机构的统一代码数据，并对采集到的数据进行预处理。通过完整的数据信息训练多种补全算法模型，并使用这些训练好的模型对缺失数据进行补全，得到多个补全值。通过随机删除完整数据中的部分属性，计算每个算法的补全精度，并将该精度作为补全值的可信度指标。通过D-S证据理论计算多个补全值的基本信任函数值，并对这些补全值进行融合，生成最终的融合结果。利用基本信任函数值对融合后的结果进行可信度量化。

2.2 多算法补全模型

2.2.1 数据预处理

数据预处理是确保补全算法有效性和结果可靠性的关键环节，主要包括异常值检测与剔除、非结构化数据转换和特征标准化。针对数值型特征进行异常值检测，依据领域知识设定合理阈值：若字段逻辑上不应为负值，如注册资金和从业人数为负值，则判定为异常记录并予以剔除。通过调用百度地图地理编码服务接口实现地址解析，将非结构化的文本地址转换为经纬度坐标。为消除量纲差异对算法的影响，对经度、纬度进行归一化处理，具体公式如式（1）所示：

$$t_i = \frac{t_i - \min(T)}{\max(T) - \min(T)} \quad T = \{t_1, t_2, t_3, \cdots, t_n\}, i \in [1, n] \qquad (1)$$

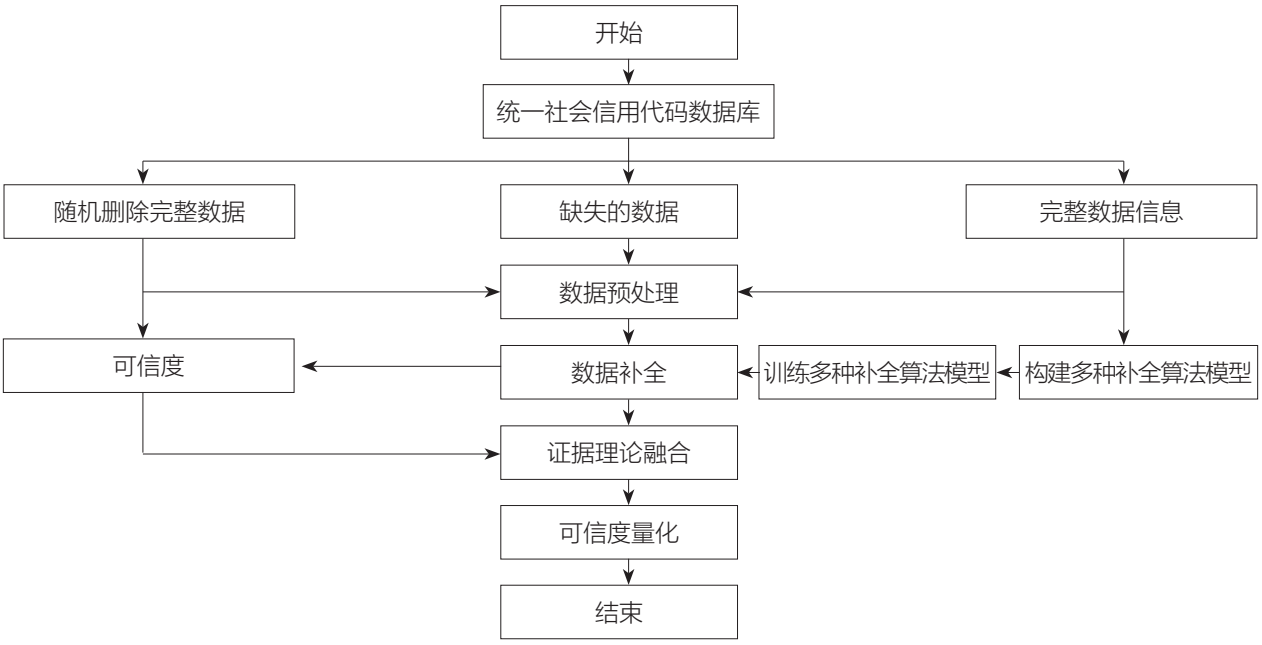

图1 D-S证据理论的统一代码缺失数据多策略融合补全方法框架图

式中：t_i表示第i个经度或纬度；T表示所有经度或纬度的集合；n为数据总量。

同时，将经度、纬度、注册资金、从业人数转化为浮点型数据。经过上述的数据预处理流程，成功构建了统一代码数据特征集，显著提升了数据的完整性和一致性，为后续多算法补全与证据理论融合奠定了坚实的数据基础。

2.2.2 多模型训练

本节构建了一个多算法融合的补全框架，结合均值填充、回归分析、K-最近邻、随机森林及NGBoost等多种算法，充分利用不同算法的优势以提升补全鲁棒性。

在模型训练阶段，将企业统一代码数据的特征集按照一定比例随机划分为训练集和测试集，以训练集和测试集中的完整特征值作为输入特征，将缺失值作为目标数据。采用均值填充、回归分析、K-最近邻、随机森林、NGBoost等多种算法，在完整数据集上训练模型。

2.3 证据理论融合

2.3.1 可信度计算

在多算法补全模型中，每个补全算法被视为独立的证据源，其输出的补全结果需通过可信度权重反映其可靠性。为量化各算法的可信度，本文提出1种基于模拟缺失实验的动态评估方法：从完整数据集中随机删除部分已知属性值，构造人工缺失数据集；调用所有补全算法对缺失值进行预测，并将预测结果与原始真实值进行比对。算法精度通过各补全算法对缺失数据进行补全，计算正确预测样本数与总删除样本数的比例，如式（2）所示：

$$w_j=\frac{Correct_j}{N_{\text{total}}}\times 100\%,\ j\in[1,t] \qquad (2)$$

式中：w_j表示第j个算法的精度；$Correct_j$为第j个算法预测正确的数量；N_{total}为总删除样本数。

2.3.2 识别框架

多算法补全模型中t种补全算法预测的补全值表示为$x=\{x_1, x_2, \ldots, x_t\}$。根据D-S证据理论设计识别框架$\theta=\{\theta_1, \theta_2, \ldots, \theta_t\}$，$\theta_i$命题表示$x_i$为正确的补全值，$\theta$是所有可能的假设全集。

2.3.3 基本概率分配

将每种不同的补全算法视为一个独立的证据源，表示为$m=\{m_1, m_2, \ldots, m_j\}$，$j\in[1,t]$，每个补全结果的基本概率分配可表示为$m_j(\{\theta_i\})=w_j$，$m_j(\theta)=1-$

w_j, $i\in[1,n]$, $j\in[1,t]$。

其中$m_i(\{\theta_i\})$的值表示证据源m_j对命题θ_i（即x_i为正确补全值）的支持程度，w_j表示在补全值为x_i的情况下，j算法对数据的可信度，$m_i(\theta)$表示证据源m_j将1–w_j的基本概率分配给不确定的假设（即全集θ）。

2.3.4 Dempster–Shafer融合规则

Dempster–Shafer合成公式通过融合多证据源的可信度，解决补全结果间的冲突并生成综合信任函数。定义冲突系数K以量化不同证据源之间的不一致性。当K=1时，表明所有证据源的补全结果完全冲突；当K=0时，表明无冲突，证据间高度一致。其计算公式如式（3）所示：

$$K=\sum_{\theta_1\cap\theta_2\cdots\theta_n=\varnothing} m_1(\theta_1).m_2(\theta_2)..m_t(\theta_n) \tag{3}$$

基于冲突系数计算补全值的合成基本信任函数值，最终得到每个命题θ_i的基本信任函数值，所述计算公式如式（4）所示：

$$M(\{\theta_i\})=\begin{cases}\dfrac{\sum_{\theta_1\cap\theta_2\cap\ldots\cap\Theta=\theta_i} m_1(\theta_1).m_2(\theta_2)..m_t(\Theta)}{1\text{-}K}, \theta_i\neq\varnothing \\ 0, \theta_i=\varnothing\end{cases} \tag{4}$$

最终的预测值 $\hat{x}$ 可以通过这些基本信任函数值加权平均来计算，所述计算公式如式（5）所示：

$$\hat{x}=\frac{\sum_{i=1}^{n} x_i\cdot M(\{\theta_i\})}{\sum_{i=1}^{n} M(\{\theta_i\})} \tag{5}$$

2.4 可信度量化

为评估多算法融合结果的不确定性，本文引入信息熵理论对基本信任函数值的分布进行量化分析。信息熵是衡量概率分布混乱程度的核心指标，其值越大表明信任函数值的分布越分散，融合结果的不确定性越高，对应的可信度越低。信息熵所述计算公式如式（6）所示：

$$H=-\sum_{i=1}^{n} M(\{\theta_i\})\log(M(\{\theta_i\})) \tag{6}$$

可信度量化值 $C(\hat{x})$ 可通过熵的倒数进行定义，即通过将信息熵取倒数的方式，将数据不确定性程度转换为可衡量的可信度指标，如式（7）所示：

$$C(\hat{x})=\frac{1}{1+H} \tag{7}$$

3　实验结果与分析

3.1 实验设置

实验数据来源于云南省统一代码数据库，涵盖机构名称、统一代码、注册地址、经济行业、注册资金、从业人数等字段。从云南省统一代码数据库中筛选无缺失记录作为基准数据集，确保数据的初始完整性。为模拟真实场景中的数据缺失问题，随机删除基准数据中20%的注册资金与从业人数字段，构建人工缺失数据集。这一设计既保留了数据的原始分布特性，又能有效评估补全算法在部分缺失条件下的性能。本实验中多算法补全模型采用K–最近邻、随机森林、NGBoost、XGBoost 4种算法。在K–最近邻模型中，邻居数设置为5，这一参数决定了在预测缺失值时参考的最近邻数据点的数量，平衡了模型的整体稳定性。随机森林模型的决策树数量设置为100，该参数反映了模型的复杂度和预测能力，通过合理的决策树数量来提高模型的稳定性和准确性。NGBoost模型的迭代次数设置为500，表示模型在训练过程中进行的优化步骤数量，较高的迭代次数有助于模型更好地拟合数据，同时避免过拟合。XGBoost模型的最大深度设置为6，用于控制决策树的复杂度，确保模型保持良好的泛化能力。其中每个模型参数设置如表3所示。

表3 模型参数

模型名称	关键参数	参数值
K–最近邻	邻居数	5
随机森林	决策树	100
NGBoost	迭代次数	500
XGBoost	最大深度	6

3.2 评价指标

为评估数据补全算法的性能，本实验将信息熵用于量化多算法融合结果的不确定性，如式（8）所示。

$$\text{Entropy} = -\sum_{H_i \subseteq \Theta} m_{\text{fused}}(H_i)\log_2 m_{\text{fused}}(H_i) \qquad (8)$$

式中：$m_{\text{fused}}(H_i)$表示对命题H_i的合成信任度；Θ为所有可能命题的集合。熵值越低，表明融合结果越集中于某一候选值，可信度越高；反之，熵值越高，则表示结果分布分散，不确定性大。

3.3 实验结果分析

为了量化各补全算法的可信度，本实验采用模拟缺失数据进行计算。从基准数据集中随机删除20%的“注册资金”与“从业人数”字段，生成人工缺失数据集。依次调用K-最近邻、随机森林、NGBoost及XGBoost算法对缺失值进行补全，并将补全结果与真实值进行对比，得出各算法的可信度，结果如表4所示。本实验将每个补全模型视为独立的证据源，利用Dempster-Shafer合成公式融合多证据源的可信度，以解决补全结果间的冲突并生成综合信任函数。定义冲突系数K，用于量化不同证据源之间的不一致性，实验得出冲突系数参数值为0.16。基于冲突系数计算补全值的合成基本信任函数值，进而得到每个命题的基本信任函数值，通过这些基本信任函数值的加权平均计算出预测值。采用该多策略融合补全方法对统一代码缺失数据进行补全后的信息熵值为0.45%。

实验结果表明，D-S证据理论驱动的统一代码缺失数据多策略融合补全方法在统一代码补全任务中展现出较好的性能，所得补全的信息熵为0.45%。实验验证了多补全算法的融合可靠性，为政府监管与企业信用评估提供了高可信度的数据修复方案。

表4 模型可信度

模型名称	可信度/%
K-最近邻	54
随机森林	62
NGBoost	27
XGBoost	41

4 结 语

本文提出了一种基于D-S证据理论的统一代码多种数据补全融合方法，旨在解决传统数据单一补全方法在处理统一代码数据时存在的局限性。通过综合运用均值、回归分析、K-最近邻、随机森林、NGBoost等多种数据补全算法，并结合证据理论对补全结果进行融合与可信度量化。本方法能够有效提升数据补全的准确性和可靠性。

在实验过程中，通过随机删除完整数据中的部分属性来模拟数据缺失场景，并计算各算法的补全精度作为可信度指标。结果表明，不同算法在处理不同类型的数据时各有优势，而通过证据理论融合后的最终补全结果，不仅充分利用了各算法的优势，还通过基本信任函数值对补全结果的不确定性进行了量化评估，进一步提高了数据的可信度。

本研究不仅为统一代码数据的补全提供了一种高效、可靠的方法，还为大数据环境下的数据补全问题提供了一种新的思路。未来将进一步优化算法融合策略，探索更多类型的补全算法，并将其应用于更广泛的数据类型和领域，以推动数据补全技术的进一步发展。同时也将关注如何进一步提高数据补全的实时性和动态适应性，以满足不断变化的市场需求和应用场景。

参考文献

[1] 全国信用标准化技术工作组.法人和其他组织统一社会信用代码编码规则:GB 32100—2015[S].北京:中国标准出版社,2015.

[2] 赵涛,李卓睿.统一社会信用代码数据服务研究[J].中国标准化,2025(3):63-67.

[3] 张根红,安鸿志,吴建军,等.统一社会信用代码在卫生监督执法领域信用监管中的应用探讨[J].中国卫生监督杂志,2021,28(3):266-270.

[4] 魏兵兵.推进统一社会信用代码工作的思考[J].经贸实践,2019(1):295-296.

[5] 齐广儒.统一社会信用代码档案数据在社会诚信建设中的作用探讨[J].陕西档案,2025(1):57-58.

[6] 赵捷,袁辉,邓祥武,等.法人和其他组织统一社会信用代码数据国民经济行业数据质量提升方法研究[J].中国标准化,2024(5):66-71.

[7] EMMANUEL T, MAUPONG T, MPOELENG D, et al. A survey on missing data in machine learning[J]. Journal of Big data, 2021(1):140.

[8] HUSSAIN S, MUSTAFA M W, AL-SHQEERAT K H A, et al. A novel feature-engineered-NGBoost machine-learning framework for fraud detection in electric power consumption data[J]. Sensors, 2021, 21(24): 8423.

[9] KARAMTI H, ALHARTHI R, ANIZI A A, et al. Improving prediction of cervical cancer using knn imputed smote features and multi-model ensemble learning approach[J]. Cancers, 2023, 15(17): 4412.

[10] LEE Y, LEITE W L. A comparison of random forest-based missing imputation methods for covariates in propensity score analysis[J]. Psychological Methods, 2024.

[11] KHAN S I, HOQUE A S M L. SICE: an improved missing data imputation technique[J]. Journal of big Data, 2020, 7(1): 37.

[12] BAI L, JI B, WANG S. SAE-Impute: imputation for single-cell data via subspace regression and auto-encoders[J]. BMC bioinformatics, 2024, 25(1): 317.

[13] LIU J, WAN Z, HU X, et al. Safe drug recommendation through forward data imputation and recurrent residual neural network[J]. Applied Soft Computing, 2024(161): 111723.

[14] LENA P D, SALA C, PRODI A, et al. Methylation data imputation performances under different representations and missingness patterns[J]. Bmc Bioinformatics, 2020(21):1-22.

[15] RAZAVI-FAR R, CHENG B, SAIF M, et al. Similarity-learning information-fusion schemes for missing data imputation[J]. Knowledge-Based Systems, 2020(187): 104805.