引用格式： 朱勋程，赵忆宁，李宁，等.信用信息治理中混合型缺失数据补全方法探索：以统一社会信用代码为例[J].标准科学，2025(10):73-78.
ZHU Xuncheng,ZHAO Yining,LI Ning,et al. Exploration of the Method for Completing Hybrid Missing Data in Credit Information Governance: A Case Study of the Unified Social Credit Code [J].Standard Science,2025(10):73-78.

信用信息治理中混合型缺失数据补全方法探索：以统一社会信用代码为例

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】随着社会信用体系建设的深入推进，法人和其他组织统一社会信用代码（以下简称为统一代码）作为信用信息治理的核心要素，其数据完整性不足问题日益凸显，制约了信用评估与监管效能。当前统一代码数据库存在混合型数据缺失问题，亟须系统性补全治理。【方法】针对混合型属性缺失问题，本文提出采用NGBoost构建概率预测模型实现混合型数据补全。【结果】构建了混合型缺失数据补全模型，提升了统一代码数据的完整性、准确性与场景适配性。【结论】研究成果为破解信用信息治理中的数据质量瓶颈提供了新的技术路径，对完善信用基础设施、挖掘数据资产价值具有重要的理论指导意义与实践应用价值。

关键词： 统一社会信用代码；数据质量提升；NGBoost
DOI编码： 10.3969/j.issn.1674-5698.2025.10.010

Exploration of the Method for Completing Hybrid Missing Data in Credit Information Governance: A Case Study of the Unified Social Credit Code

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

Abstract: [Objective] With the in-depth promotion of the construction of the social credit system, the unified social credit code becomes the core element of credit information governance. However, the lack of integrity of the unified social credit code data emerges, which has constrained the effectiveness of credit assessment and supervision. Currently, the unified social

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，学士，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，学士，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。

credit code database has the problem of mixed-type data missing, so it is urgently necessary to conduct systematic completion and governance. [Methods] Regarding the problem of mixed-type attribute missing, this paper proposes to use NGBoost to build a probability prediction model to achieve mixed-type data completion. [Results] This paper has constructed a mixed-type missing data completion model, so as to improve the completeness, accuracy and scene adaptability of the unified social credit code data. [Conclusion] The research results provide a new technical path for solving the data quality bottleneck in credit information governance, and have important theoretical guiding significance and practical application value for improving the credit infrastructure and extracting the value of data assets.

Keywords: unified social credit code; data quality improvement; NGBoost

0　引言

信用信息治理作为现代信息化社会体系构建的重要基石，通过系统化整合政府监管、金融风控、市场交易等多源数据，构建了覆盖全社会的信用评估与协同管理网络。其核心目标在于打破信息孤岛，实现跨部门、跨领域的数据互联互通，为公共决策、风险预警、资源配置提供精准支撑。在技术驱动下，信用信息治理逐步形成“数据采集—标准化处理—智能分析—场景应用”的全链条闭环，不仅提升了市场监管效能，还通过信用画像、联合奖惩等机制推动经营主体的协同发展，成为优化营商环境、防范系统性风险的关键基础设施。

统一社会信用代码（Unified Social Credit Code，USCC）是我国法人及其他组织的唯一法定身份标识。2015年6月11日，国务院发布《国务院关于批转发展改革委等部门法人和其他组织统一社会信用代码制度建设总体方案的通知》（国发〔2015〕33号）[1]，标志着我国统一代码制度全面实施。USCC作为信用信息体系的核心数据枢纽，通过“一码关联”打通工商登记、税务申报、社保缴纳等多维数据，为构建全景式信用档案提供了基准索引。高质量的统一代码数据不仅为宏观经济分析提供可靠依据，更在微观层面赋能企业信用评估、供应链风险管理等场景，推动数据价值进一步挖掘。

1　概述

统一代码是信用信息治理体系的核心要素，是国家推进商事制度改革、优化营商环境的重要制度设计。该代码以唯一性、兼容性、稳定性及全覆盖性为核心特征：唯一性体现为每个法人和其他组织自设立时即获得终身不变的法定数字身份标识，有效解决传统多码并存导致的数据孤岛问题；兼容性则通过科学设计的18位代码结构得以体现，代码整合了原无含义的9位组织机构代码，增加有含义的代码位，满足各部门的需求，实现跨部门编码体系的系统衔接；稳定性体现为统一代码在全生命周期内不因主体信息变更而调整，形成可追溯的管理闭环；全覆盖性体现为统一代码突破行业、地域及组织形式的限制，将企业、个体工商户、农民专业合作社等经营主体全部纳入统一编码体系。

统一代码是社会信用体系的重要组成部分，提升统一代码数据质量，可更好地促进社会诚信建设[2]。作为信用信息的基准索引，统一代码通过结构化编码规则承载多维主体特征信息：前段登记管理部门代码精准映射注册部门，中段行政区划代码反映注册地溯源信息，后段主体标识码构建唯一性保障机制。这种标准化编码机制显著提升了政府治理效能，既避免了多部门重复采集信息造成的行政资源浪费，又降低了企业填报频次，切实减轻经营主体负担。

在跨部门协同层面，依托统一代码构建的“一码关联”机制，实现了市场监管、税务、金融等40余个部门间的数据实时共享与业务联动，极大地促进了信用信息治理体系的建设。

在应用价值层面，统一代码已深度融入经济社会治理的各个环节。统一代码数据能够有效破解业

务库所面临的数据孤岛难题[3]，在政务管理[4]、信用监管[5]、数字经济[6]、身份认证[7]等社会管理和经济活动的多个领域，充分发挥了实名管理与分析决策的关键作用，为各领域的高效运转和协同发展提供了有力支撑。

2 问题分析

我国大量统一代码数据存在属性缺失问题，问题的成因涉及历史遗留、制度协同及技术应用等多重维度[8]，亟待系统性补全治理。

（1）早期数据采集不规范，导致问题数据不断积累。早期存量数据因采集标准缺失、录入方式粗放导致质量缺陷，1990—2010年，经营主体登记信息依赖人工转录纸质档案，字段完整率不足，部分关键属性（如出资人信息、经营范围编码）存在系统性漏录。

（2）跨部门数据协同治理措施失效，没有将多部门数据有效整合，从而引发数据缺失。部门间数据生产标准尚未全面统一，市场监管、税务、海关等核心业务系统存在字段定义差异，导致企业行政许可信息与统一代码无法有效关联匹配；数据更新机制缺乏联动，变更信息在部门间的传递存在时差，形成监管盲区。

（3）基层数据采集能力薄弱导致数据持续性缺失。受财政投入与技术水平制约，统一代码制度实施初期，部分机构采用离线填报系统，无法实时对接信用信息平台实施校验，数据及时性和完整性存在差距；数据协同校验机制不完善，导致错误或缺少属性的数据快速累积。

上述数据缺失问题已对信用信息治理体系形成实质性制约。例如，金融领域因代码关联信息不全导致信贷风险评估偏差，市场监管部门因数据断层造成异常经营主体漏检。因此，需要构建数据修复机制，实现属性缺失数据的精准补全。

现阶段，已有针对统一代码补全的相关研究，赵捷等[9]提出对统一代码数据质量问题开展精准鉴别工作，随后构建混合型自动化行业分类模型，补全统一代码行业分类属性。但该方法的应用范围相对较窄，仅针对行业分类这一单一属性，而目前统一代码所面临的属性缺失问题，往往是多属性、混合型的复杂情况。实际上，当前已出现了多种混合型数据补全方法[10]。例如，Kowarik等[11]在2001年提出的K最近邻填补算法（K-Nearest Neighbors Imputation, KNNimpute），其基本原理是通过寻找与观测值距离最近的K个变量，再对这K个变量取加权平均值，从而得到用于填补缺失变量的值。不过，KNN方法依赖距离度量（如欧氏距离）这一特性，虽然在低维连续数据的处理场景中性能相对稳定，但对混合型数据里离散变量的处理效果却不理想，且邻居数K值需要进行精细调优。此外，Stekhoven等[12]提出了一种基于随机森林迭代机制的缺失森林（Missing Value Imputation using Random Forest, MissForest）填补算法。这一算法凭借随机森林的非参数特性和对混合数据类型的良好兼容性，得到了较为广泛的应用。然而，其迭代训练机制也导致了计算复杂度的显著增加。Audigier等[13]提出了一种因子分析（Factorial analysis for mixed data, FAMD）填补算法，算法基于主成分分析法探究个体间的相似性以及变量间的关系。但该算法基于线性关系的假设难以适应复杂非线性关联或非随机缺失模式。

为推进社会信用体系建设，解决统一代码数据库因历史数据录入不完整、多源信息分散及跨部门协同不足导致的属性缺失问题，需构建对统一代码数据进行全面补全的方法，基于多维特征关联分析开发预测模型，且模型对混合数据类型具有良好兼容性，能够适应属性复杂的非线性关联模式，自动补全缺失。该方法的建立旨在通过系统性补全机制，提升数据完整性、准确性及可用性，为政务管理、风险预警及资源优化提供可靠支撑。

3 基于NGBoost混合型数据补全方法

3.1 数据处理

本文将统一代码数据分割为属性完整的数据

集和属性缺失的数据集，记为第一数据集和第二数据集，属性总数为12，统一代码数据包含部分属性如表1所示。

表1 统一代码数据部分属性表

序号	离散型属性	连续型属性
1	行政区划	注册资金
2	经营范围	从业人数
3	企业类型	邮政编码
…	…	…

对离散型属性，采用数字编码的方式进行处理，以确保数据的一致性。以企业类型的属性值为例，其有6种属性值，分别是有限责任公司、个人独资企业、合伙企业、全民所有制企业、集体所有制企业和农民专业合作社。对企业类型的属性值进行数字编码的结果如表2所示。

表2 企业类型的数字编码表

属性值	数字编码
有限责任公司	1
个人独资企业	2
合伙企业	3
全民所有制企业	4
集体所有制企业	5
农民专业合作社	6

对连续型属性进行归一化处理，以便将不同量纲的数据转换为相同的标准。归一化处理如公式（1）所示。

$$x_{\text{normalized}} = \frac{x - \min(X)}{\max(X) - \min(X)} \tag{1}$$

3.2 模型训练

在模型训练阶段，本文对第一数据集中的属性数据进行多次不放回的随机抽取。每次抽取生成2个新的数据集：数据集A包含被抽取的属性$\{a_1, \cdots, a_m\}$，$m<12$，而数据集B则由未被抽取的属性组成$\{b_1, \cdots, b_l\}$，$l<12-m$。数据集B将作为特征属性，而数据集A中的各个属性则作为目标属性。根据目标属性的类型，选择相应的回归或分类NGBoost模型进行训练。若目标属性为连续型，则使用回归模型；若为离散型，则使用分类模型。模型训练时的特征属性和目标属性表示如公式（2）所示。

$$\begin{aligned} NGBoost_1 &\Rightarrow \mathrm{x}{:}b_1, b_2, \ldots, b_l \quad y{:}a_1 \\ &\vdots \\ NGBoost_m &\Rightarrow \mathrm{x}{:}b_1, b_2, \ldots, b_l \quad y{:}a_m \end{aligned} \tag{2}$$

在所有缺失属性的预测模型训练完成后，将所有NGBoost模型组合，形成一个复合NGBoost模型，以便后续对缺失数据进行预测。

3.3 缺失属性的补全

在进行缺失属性补全时，首先利用构建好的复合NGBoost模型对第二数据集中缺失属性的概率分布进行预测，即模型将根据已知属性推断缺失属性的可能值，输出各个预测结果的概率密度函数值或概率值。具体的预测公式为：

对于连续型属性：

$P(Y=y|X)=f(y)$

对于离散型属性：

$P(Y=y|X)=P(y|x)$

补全过程将设定一个合理的阈值，以提高补全结果的可靠性。当预测的概率分布最大值达到设定的阈值时，才能进行缺失属性的补全。若缺失属性为连续型，则采用归一化的逆变换方法；若为离散型属性，则使用解码的方法进行补全。相应的补全公式为：

对于连续型属性：

$$\hat{y} = (\max(Y_j) - \min(Y_j)) \times y_{p\max} + \min(Y_j)$$

对于离散型属性：

$$\hat{y} = Decode(y_{p\max})$$

通过以上3个步骤的实施，基于NGBoost的混合型数据补全方法能够高效、准确地对统一代码数据中的缺失属性进行补全，显著提升数据的完整性与可靠性，为后续的数据分析和挖掘提供坚实的基础。

4 实验结果与分析

实验抽取注册资金为目标属性，经度、纬度、

从业人数、经济行业作为特征属性，选择回归模型进行训练，概率密度函数值的阈值设定为0.35。为了构建出最合适的NGBoost模型，本文通过实验分析算法模型的参数：学习率、迭代次数。为了全面评估各参数对模型性能的影响，选择均方根误差（Root Mean Square Error，RMSE）作为实验的评价指标。RMSE公式如公式（3）所示。

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (3)$$

学习率：控制模型每次迭代的步长，学习率越小，模型的训练速度越慢；学习率越大，则越可能导致模型过拟合。不同学习率的NGBoost模型实验表现分别如图1所示（此时的迭代次数=500）。

从图1中可以看出，当学习率上升至0.07时，预测注册资金的精度达到最佳值。

迭代次数：指决策树的数量。该参数可以提高模型的准确性，但也会增加模型的计算时间。不同迭代次数的实验表现如图2所示（此时的学习率=0.5）。

从图2中可以看出，当迭代次数增加至500时，预测注册资金的精确度基本趋于稳定。因此，预测注册资金的NGBoost模型确定学习率为0.07，迭代次数为500。

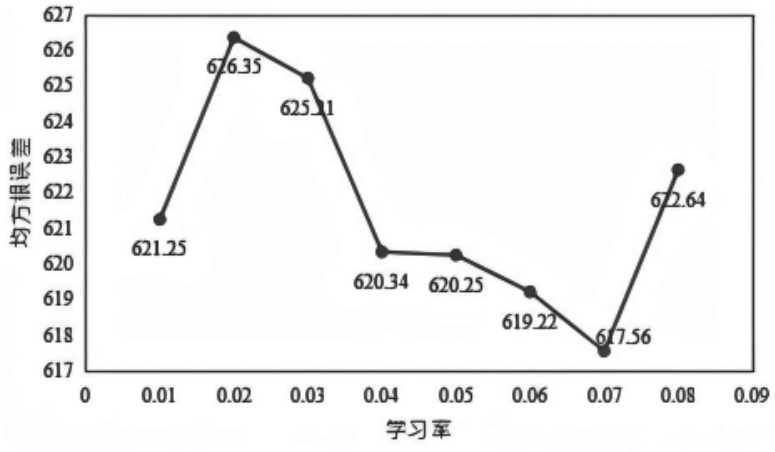

图1 不同学习率下的均方误差折线图

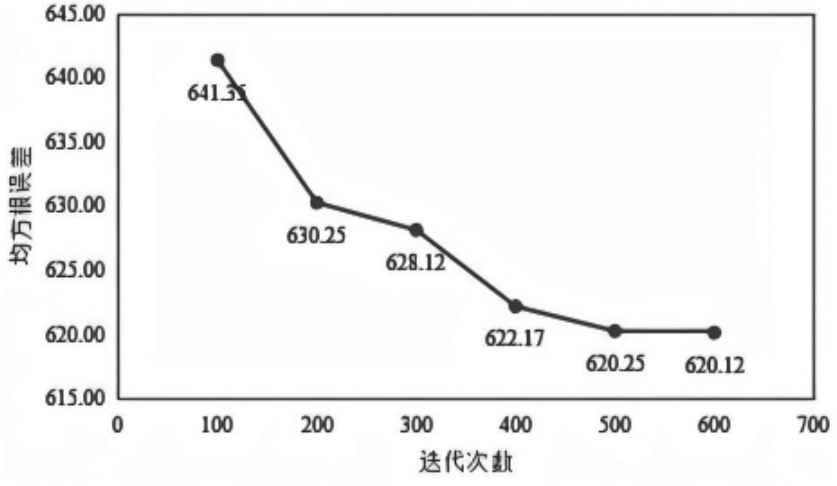

图2 不同迭代次数下的均方误差折线图

5 结 论

本文分析了信用信息治理的内涵特征与发展态势，探索了以统一代码为核心要素的治理机制，同时阐释了统一代码的结构特征及其在信用信息治理体系中的功能定位，并揭示了当前数据质量存在的突出问题，进而构建了基于NGBoost的补全方法框架。研究成果对提升统一代码数据质量、深挖数据资产价值具有重要的理论指导意义与实践应用价值。

当前，我国正处于信息化建设的关键阶段，机器学习技术迅猛发展，并深度融入国家信息化体系构建进程，为信用信息治理体系建设提供了坚实的技术支撑。作为国家治理现代化的重要标志，信用信息治理体系的完善程度直接关系社会信用生态的构建质量。其中，统一代码作为基础性数据要素，其数据完整性与质量水平直接影响信用信息治理效能。通过创新性的数据补全技术实现代码信息的精准补全与质量提升，不仅能有效强化信用主体身份标识的精准性，更可帮助深度挖掘统一代码数据资产的应用价值，为构建智慧化信用治理体系奠定数据基石。

参考文献

[1] 国务院.国务院关于批转发展改革委等部门法人和其他组织统一社会信用代码制度建设总体方案的通知[EB/OL].(2015-06-17)[2024-11-28]. https://www.gov.cn/zhengce/zhengceku/2015-06/17/content_9858.htm.

[2] 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.

[3] 周烨.法人及其他组织统一社会信用代码数据在大市场监管信息化工作中的应用研究[J].科技风,2019(26):265-266.

[4] 刘吉洲,张永全,郑伟,等.区域性统一社会信用代码信息服务实践与研究:以山东省济宁市为例[J].中国标准化,2020(7):114-118.

[5] 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.

[6] 周顺骥.基于福建省法人和其他组织统一社会信用代码的数字经济发展状况研究[J].中国质量与标准导报,2022(6):75-79.

[7] 黄润飞,陈贤明,黄燕玲,等.基于身份标识和区块链技术的粤港澳大湾区法人及其他组织跨境身份认证应用研究[J].标准科学,2023(8):53-57.

[8] 王秀峰,杨德富,蔡欣畅.信用监管体系下统一社会信用代码数据质量提升方法研究[J].中国标准化,2025(4):33-39.

[9] 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.

[10] 杨弘,田晶,王可,等.混合型缺失数据填补方法比较与应用[J].中国卫生统计,2020,37(3):395-399.

[11] KOWARIK A,TEMPL M.Imputation with the R Package VIM[J].Journal of Statistical Software,2016,74(7):1-16.

[12] STEKHOVEN D J, BÜHLMANN P. MissForest—nonparametric missing value imputation for mixed-type data[J]. Bioinformatics, 2012, 28(1): 112-118.

[13] AUDIGIER V, HUSSON F, JOSSE J.A principal component method to impute missing values for mixed data[J].Advances in DataAnalysisand Classification,2016,10(1):5-26.