Advances in Applied Mathematics 应用数学进展, 2025, 14(10), 420-425
Published Online October 2025 in Hans. https://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2025.1410453

Hans汉斯

数据要素流通中大数据分析的数学方法及实践探索

刘　冲[1]，张　瑶[2]

[1]河北金融学院金融科技学院，河北　保定
[2]河北金融学院管理学院，河北　保定

收稿日期：2025年9月28日；录用日期：2025年10月22日；发布日期：2025年10月29日

摘　要

本文针对数据要素流通当中所含的大数据分析的数学建构和利用情况，先阐述支持数据要素流通的关键技术架构，重点剖析隐私算账，区块链等关键技能；接着梳理统计学，线性代数，改善理论这些核心数学工具，探讨它们在数据预处理，特征提取，创建模型中的具体应用情形；再全面考量数据品质，算法公正，信息安全，场景契合这些实际问题；最后从数据治理，算法优化，安全保障等很多层面给予改善建议。期望给提升数据要素流转效能，推进数据价值展现赋予理论依据和执行参照。

关键词

数据要素流通，大数据分析，数学方法，应用

Mathematical Methods and Practical Exploration of Big Data Analysis in the Circulation of Data Elements

Chong Liu[1], Yao Zhang[2]

[1]School of Financial Technology, Hebei Finance University, Baoding Hebei
[2]School of Management, Hebei Finance University, Baoding Hebei

Received: September 28, 2025; accepted: October 22, 2025; published: October 29, 2025

Abstract

This article focuses on the mathematical construction and utilization of big data analysis contained in the circulation of data elements. It first elaborates on the key technical architecture that supports

文章引用: 刘冲, 张瑶. 数据要素流通中大数据分析的数学方法及实践探索[J]. 应用数学进展, 2025, 14(10): 420-425.
DOI: 10.12677/aam.2025.1410453

the circulation of data elements, with a particular emphasis on analyzing key technologies such as privacy accounting and blockchain. Next, sort out the core mathematical tools such as statistics, linear algebra, and improvement theory, and explore their specific application scenarios in data preprocessing, feature extraction, and model creation; comprehensively consider the actual issues such as data quality, algorithm fairness, information security, and scenario fit. Finally, improvement suggestions are given from many aspects such as data governance, algorithm optimization, and security guarantee. It is expected to provide theoretical basis and implementation reference for enhancing the efficiency of data element circulation and promoting the demonstration of data value.

Keywords

Circulation of Data Elements, Big Data Analysis, Mathematical Method, Application

1. 引言

数字技术飞速发展的时代背景下，数据规模出现爆发式增长态势，中国信息通信研究院给出的统计预估显示，到 2024 年全球数据总量大概会达到 180 亿字节(ZB)，这显示出数据要素的战略重要性，数据要素是数字经济的关键推动因素，其高效流通可以消除数据孤岛，促使不同行业整合，做到跨区域协同共享，从而推动资源得到优化配置，目前，数据要素流通碰到体量庞大，种类繁多，价值密度低，传输效率不够高等诸多难题，迫切需要依靠大数据分析技术来改进处理效能和价值发掘水平。

大数据分析技术依靠海量、多源、异构的数据资源，利用数学工具来挖掘数据中的信息价值，数学理论是其重要支撑，给数据预处理、特征提取、模型搭建以及结果验证这些环节赋予了理论根基，从基础的统计学知识到复杂的机器学习算法，数学方法渗透在数据要素流动的各个环节当中，对提高分析的精确度和可信度有着决定性作用，仔细探究数据要素流通中的数学方法及其实际应用价值，对于健全数据要素市场体系，最大程度地发挥数据要素的效益具有重要的现实意义[1]。

2. 数据要素流通的关键技术

2.1. 隐私计算技术

想要做到数据要素的安全流通，关键之处就在于利用联邦学习，多方安全计算等这些前沿的技术手段，在保证数据隐私的情况下进行分析挖掘工作，从而达到“可用不可见”的效果，这种技术既能避免数据传输过程中出现泄露的风险，又能够很好地满足实际应用当中对于数据分析的要求。

2.2. 区块链技术

区块链技术在数据流通领域应用有着明显的价值优势，它的去中心化特性能很好地解决数据权属界定不清的问题，而且不可篡改的特性保证了数据的真实性和可靠性，利用区块链技术，数据从产生到应用的整个过程都能做到透明追溯，给形成科学规范的数据产权体系给予了可靠的科技支持[2]。

2.3. 数据空间技术

数据空间作为依靠共识机制来开展多方协作的平台，它成了支撑数据流通和应用的关键基础设施，

其关键技术架构包含很多方面，像数据资源的供应管理，传输监管以及开发利用等等，而且通过双向接口促使各个参与方加入到网络环境之中，从而做到数据跨域整合并且得到有效地利用[3]。

2.4. 人工智能技术

人工智能大模型技术发展起来以后，就给数据要素高效流转赋予了新的解决办法，凭借智能算法，可以做到对数据进行细致划分和分层管理，从而改善数据处理的安全性以及运行速度，而且通过深入挖掘和解析的办法，能够体现数据隐私的价值和应用情形，此技术还能改良数据流通过程中的算法规划和模型搭建步骤，进而加强数据传送的稳定性和精确度(表 1) [4]。

Table 1. Application effect comparison of key technologies in data element circulation
表 1. 数据要素流通关键技术应用成效对比

关键技术	应用场景	数据泄露风险降低率	数据处理效率提升率	跨域传输速率提升率	典型案例成效
隐私计算(联邦学习)	金融客户信用评估	82%	35%	-	某银行联合建模后坏账预测误差降低 12%
区块链技术	政务数据跨部门共享	95%	28%	42%	某省政务平台数据纠纷量下降 68%
数据空间技术	制造业供应链数据整合	78%	45%	53%	某车企供应链响应时间缩短 40%
人工智能大模型	电商用户行为分析	65%	60%	30%	某电商推荐转化率提升 25%

3. 数据要素流通中大数据分析的核心数学方法

3.1. 数据预处理：解决流通数据的“异构性”问题

数据要素流通的首要挑战是数据格式不统一、存在噪声与缺失值，需通过数学方法标准化处理[5] [6]。

数据归一化：消除量纲差异对分析结果的影响，常用 Min-Max 归一化公式：其中，x 为原始数据，min(x)、max(x)分别为数据集中的最小值与最大值，为归一化后的数据(取值范围[0, 1])。

缺失值填充：采用 K 近邻(KNN)插值法，通过计算样本与周边 K 个样本的欧氏距离加权填充，欧氏距离公式为：

$$d\left(x_i, x_j\right)=\sqrt{\sum_{k=1}^{n}\left(x_{ik}-x_{jk}\right)^2} \tag{1}$$

其中，x_i, x_j 为两个样本，n 为特征维度，x_{ik} 、x_{jk} 分别为样本 i、j 在第 k 个特征上的取值。

3.2. 关联规则挖掘：挖掘流通数据的“隐性价值”

在数据要素流通场景中，需识别不同主体、不同场景数据的关联规律，Apriori 算法是经典数学工具，其核心是基于“频繁项集支持度阈值”筛选有效关联规则。

支持度计算：项集 X 在数据集 D 中的支持度表示 X 出现的频率，公式为：

$$\sup port\left(X\right)=\frac{count\left(X\subseteq T, T\in D\right)}{\left|D\right|} \tag{2}$$

其中，$count\left(X\subseteq T\right)$ 为包含项集 X 的事务数，$\left|D\right|$ 为数据集总事务数。

置信度计算：关联规则 $X\rightarrow Y$ 的置信度表示包含 X 的事务中同时包含 Y 的概率，公式为：

$$confidence\left(X\rightarrow Y\right)=\frac{\sup port\left(X\cup Y\right)}{\sup port\left(X\right)} \tag{3}$$

3.3. 隐私保护计算：保障流通数据的“安全性”

数据要素流通需平衡价值挖掘与隐私保护，差分隐私是主流数学方法，通过添加噪声实现数据匿名化。其核心是 ϵ-差分隐私定义：若对于任意两个仅相差一条记录的数据集 D_1, D_2，以及任意输出集合 S，满足：

$$\Pr\left[M\left(D_1\right) \in S\right] \leq e^{\epsilon} \cdot \Pr\left[M\left(D_2\right) \in S\right] \tag{4}$$

其中，M 为隐私保护算法，ϵ 为隐私预算(ϵ 越小，隐私保护强度越高)。

4. 数据要素流通中大数据分析面临的挑战

4.1. 数据质量参差不齐，影响数学方法的应用效果

数据质量作为大数据分析的一项关键支持要素，重要性不言而喻，数据要素在各个领域的跨界流转过程中，因其不同的数据源头、缺少标准化的现状，经常伴随产生诸如缺失值、异常值、重复数据等一系列质量问题。一方面一些数据提供者为了谋求规模效益而轻视预处理阶段的任务，造成原始数据偏差较大且不够完整，另一方面由于不同机构之间格式、编码方式存在差异性特点，在合并多类型异质数据时就容易出现兼容性方面的难题，这些情况都直接影响到各类统计方法的应用表现情况：比如在回归操作里头由于存在过高的缺失频率使得该模型中的参数估计可能产生偏倚；又或者当涉及到基于分类原理设计出特定算法的时候异常值现象也会令最终得到的划分过程变得不稳定——如此一来整体结论可靠度也随之下降了。

4.2. 算法公平性不足，存在偏见与歧视风险

在金融信贷、人力资源调配以及行政管理这些领域的数字化转型进程中，大数据技术的大量应用促使人们开始重视起算法是否公平的问题，有些数学建模手段由于训练样本带有偏差或者算法本身的设计存在缺陷，就有可能造成预测结果发生系统性的失真状况，在信贷风险管理这一情形之下，如果过往的数据大多显示某个群体有着负面的标签表现，那么逻辑回归或者深度神经网络所搭建起来的模型大概会对这个群体存在歧视倾向，进而大幅降低该群体申请贷款的通过率，在招聘筛选环节当中，假如招聘平台训练用的数据包含了性别或者年龄方面的歧视信息，那么推荐结果就会偏向那些具有特定属性特征的求职者，从而违反了公平的原则，算法偏见不但损害到个体的权益，而且可能妨碍整个社会公正性目标的达成，给数据要素市场的发展带来一定的限制作用。

4.3. 数据安全与隐私保护压力大，制约数据流通范围

数据安全和隐私保护是数据要素流通的基本前提条件，数据在收集、存储与处理各个阶段的大数据分析中都潜藏着安全问题，首先大量数据在传输过程中容易受到攻击或泄露，侵犯了个人权益；其次有些算法的运行需要借助用户敏感信息才能正常完成相应功能，如果没有做好严格的防护工作就极有可能造成隐私泄漏事故，在医疗方面，如果病人信息没有经过脱敏而被应用到机器学习建模中，会直接暴露患者的身体状况；在电商方面，如果买家的交易记录以及支付信息被盗取，会导致财产方面的损失。现行数据安全法律体系漏洞依旧很明显，责任划分不明，数据安全和隐私保护就陷入困境，这种不确定性让一些人担心会遇到麻烦，于是选择保守态度，这严重影响了数据流通的数量和质量。

5. 优化路径

5.1. 完善数据治理体系，提升数据质量

建立系统的数据治理体系是提高数据质量的关键路径，要创建统一的数据标准和规范体系，包含数

据格式，编码规则以及元数据管理等内容，从而达成数据的标准化采集和存储，还要形成起全生命周期的数据质量管控机制，在数据采集阶段设立校验规则，去掉无用信息，在预处理阶段利用先进算法去解决缺失值和异常值问题，在流通环节规划量化评价模型，定时展开数据质量检测，并不断改善，而且要促使数据治理朝着智能化方向迈进，依靠人工智能技术塑造智能治理平台，做到自动化数据清洗，融合和转变，极大改进治理效能和精准度。

5.2. 强化算法公平性设计，保障社会公平

要提升算法的公平性，其根本途径在于创建多方面协同优化的框架，包含数据，模型以及监管执行这三个层面，在数据这一层面，要完善样本选取以及预处理流程，清除存在偏差因素的样本，形成种类繁多又相对平衡的数据集，在搭建模型的时候，应当使用结合公平性约束的目标函数设计法，正则项控制偏见扩展，对抗学习削弱模型固有偏斜趋向，在监管方面，则要完善算法公平公正评定体系，塑造统一标准和检测手段，对关键行业的算法运用做到全程监视，加快推动算法透明化进程，促使公司公布其算法结构，所用训练数据及其做出决策的依据，从而加强公众信心并切实保护用户权利[7]。

5.3. 构建全方位安全保障体系，兼顾数据安全与流通

数据要素流通的关键问题就是怎样平衡安全保护和高效传输之间的矛盾，迫切需要形成起覆盖整个链条的安全保障体系，就技术方面而言，要着重解决加密算法，数据脱敏以及匿名化处理这些关键技术难题，切实保护好个人隐私权益，依靠区块链技术做到数据交换过程的全程追踪，从而有效地阻止篡改和滥用情况的发生，从制度层面来说，要完善有关法律法规，明确数据的所有权归属，市场准入门槛以及主体权责划分等内容，给数据有序流动给予法治支撑，而且还要推广数据沙盒，联邦学习等新的应用形式，遵照“数据可用不可见”这一理念，在保证信息安全的前提下推动资源共享和协同创新，也要强化跨部门协作和区域联动机制的创建，进一步提升综合治理能力，还要严厉打击数据领域的违法违规行为。

6. 总结

本研究重点放在数据要素流通中的大数据分析数学方法上，先全面阐述隐私计算，区块链这些关键技术如何支持数据流通，再细致剖析统计学，线性代数，优化理论等关键数学工具的原理和实际意义，文章接着探讨数据质量保障，算法公平性等现实问题，并给出解决办法，研究显示，数学方法是大数据分析的关键支撑，对于改善数据流通效率和价值达成有重大意义，展望将来，要促使数学理论和前沿技术更好地融合，创建起更为完善的综合体系，从而更好地适应数字经济时代数据要素高效流动的需求。

基金项目

河北省金融科技应用重点实验室科研基金项目，智能化银行数字员工构建及关键技术研究，立项编号：2024007。

参考文献

[1] 夏义堃, 管茜, 李纲. 数据信托的内涵, 生成逻辑与实现路径——基于数据流通视角的分析[J]. 图书情报知识, 2022, 39(5): 109-119.

[2] 王晓庆, 孙战伟, 吴军红, 等. 基于数据要素流通视角的数据溯源研究进展[J]. 现代图书情报技术, 2022, 6(1): 43-54.

[3] 刘业政, 宗兰芳, 金斗, 等. 数据要素流通使用的安全风险分析及应对策略[J]. 大数据, 2023, 9(2): 79-98.

[4] 梁伟亮. 人工智能大模型训练中数据的赋能型治理[J]. 学习与探索, 2025(3): 73-84.

[5] 刘芳. 商贸流通业发展对全要素生产率的影响实证研究[J]. 商业经济研究, 2018(11): 8-10.

[6] 周向红, 姚轶力, 刘雨欣. 数据要素流动背景下城市治理关键节点识别及影响因素分析——以上海市两区 51 个部门的数据为例[J]. 东南学术, 2023(1): 137-149.

[7] 田雪晴, 廖子锐, 邱英鹏, 等. 基于 PEST 分析的医疗健康数据要素价值释放路径研究——以深圳市实践为例[J]. 医学信息学杂志, 2025, 46(3): 1-7.