大数据环境下敏感数据识别分类分级技术优化与安全管控策略

邹婷
南京高泰信息科技有限公司，江苏 南京 210046
DOI: 10.61369/TACS.2026040015

摘　　要： 大数据环境下，敏感数据识别、分类分级与安全管控已成为数据安全治理的核心议题。针对海量异构数据带来的识别精度不足、分类标准不统一、管控策略难以落地等现实困境，本文系统分析了当前敏感数据治理的核心痛点，提出“规则 +AI”融合的智能识别技术、“业务 – 安全”双维分类分级框架及数据资产盘点与标签化管理策略，构建涵盖数据全生命周期的安全管控体系，旨在实现安全管控与业务发展的动态平衡。

关 键 词： 大数据；敏感数据；数据识别；数据分类分级；安全管控

Optimization of Sensitive Data Identification, Classification and Grading Technology and Security Control Strategies in the Big Data Environment

Zou Ting
Nanjing Gaotai Information Technology Co., Ltd., Nanjing, Jiangsu 210046

Abstract: In the big data environment, the identification, classification, grading and security control of sensitive data have become core issues in data security governance. Aiming at the practical dilemmas brought by massive heterogeneous data, such as insufficient identification accuracy, inconsistent classification standards, and difficulty in implementing control strategies, this paper systematically analyzes the core pain points of current sensitive data governance. It proposes an intelligent identification technology integrating "rules + AI", a "business-security" dual-dimensional classification and grading framework, and strategies for data asset inventory and label-based management. A security control system covering the entire data lifecycle is constructed to achieve the dynamic balance between security control and business development.

Keywords: big data; sensitive data; data identification; data classification and grading; security control

引言

随着大数据技术的快速发展和广泛应用，数据已成为关键的生产要素和战略资源。然而，数据价值的释放与数据安全风险的加剧如影随形。敏感数据的泄露、滥用和违规流转不仅给个人隐私带来严重威胁，也给企业声誉和国家信息安全造成重大损失[1]。近年来，《数据安全法》《个人信息保护法》等法律法规的相继出台，对敏感数据的识别、分类分级和安全管控提出了明确要求[2]。如何在海量数据中精准识别敏感信息、科学实施分类分级、动态配置安全管控策略，成为数据安全领域亟待突破的关键问题。

一、大数据环境下敏感数据治理核心痛点

（一）数据特性与治理能力不匹配

大数据体量巨大、类型多样、产生高速、价值密度低，与治理能力矛盾突出。数据规模指数级增长，传统识别方式效率成本不可行；数据结构复杂，多模态数据对识别技术适应性要求高；数据快速流转和动态更新，静态规则难以覆盖全貌，漏报误报普遍，导致敏感数据“看不清、认不全、管不住”[3]。

（二）识别分类分级体系不健全

许多组织在敏感数据识别与分类分级方面缺乏系统化、标准化框架。识别维度上，依赖单一技术手段，对深层特征挖掘不足，难以应对规避手段；分类标准上，不同行业部门差异大，跨系统数据共享交换面临标准冲突；分级逻辑上，简单挂钩保密等级与敏感程度，缺乏动态调整机制，结果可解释性不足，影响体系落地。

（三）安全管控与业务发展失衡

敏感数据安全管控与业务发展需求矛盾突出。部分组织采取“一刀切”强管控策略，阻碍数据价值释放；另一些组织投入不足，敏感数据在多场景缺乏有效防护，存在泄漏滥用风险。需建立差异化管控策略，实现安全与发展的动态平衡[4]。

二、大数据环境下敏感数据识别分类分级技术优化路径

（一）构建“规则 +AI”融合的智能识别体系

针对大数据环境下敏感数据识别面临的精度与效率双重挑战，构建“规则 +AI”融合的智能识别体系是可行的技术优化路径。该体系的核心设计思想是充分发挥规则引擎的确定性与人工智能的泛化能力，形成互补优势[5]。具体而言，规则引擎层基于正则表达式、关键字词典、数据格式校验等传统技术，对身份证号、银行卡号、手机号码等具有明确模式特征的敏感数据进行快速匹配，实现低延迟、高精度的初步筛选。人工智能层则采用自然语言处理、深度学习和预训练大模型技术，对规则引擎难以覆盖的非结构化文本、语义隐含信息、图像中的敏感文字等进行智能识别。在实际部署中，可构建多模型协同架构：利用命名实体识别模型抽取文本中的敏感实体，采用文本分类模型判断文档整体敏感等级，借助图神经网络挖掘数据之间的关联敏感度。为提升识别体系的持续适应能力，应引入主动学习机制，将人工复核结果作为反馈信号对模型进行增量训练，使识别能力随业务数据演化而动态提升。

（二）建立“业务 – 安全”双维分类分级框架

传统按安全保护角度对数据进行等级划分的方法主要是依据窃取该类数据可能造成的损失程度来进行等级划分，这种单一维度难以全面反映不同类别数据在各类商业活动场景中所面临的差异化风险特征。“业务 – 安全”双维层级体系是开展精细化管控的重要基础，其包含两个核心组成部分：一是从业务视角，在对数据的行业归属、对象类别、用途属性等方面进行划分，例如可将数据归属于诸如客户信息、产品数据、业务数据、财务数据、科研数据等诸多类别，在每个类别下再细分若干子类；二是在安全性方面根据数据的重要程度及危害等级进行分级定义：一般分为四层——关键、重要、一般、公共。这两者的组合构成了分类分级矩阵，为进行差异化管控提供框架指引[6]。

（三）优化数据资产盘点与标签化管理

首先应进行数据资产盘点，这是分类定级的前提条件；其次，标签化管理是将分类定级成果落地的关键环节。在推进数据资产管理过程中，应变“填表式”为“挖掘式”，借助自动扫描技术和数据地图实现：全面测绘组织内部的数据存储位置、大小、类型、流向等信息，建立动态更新的数字资产清单；对于结构化数据可通过解析元数据以及样本抽样进行特征推断的方式实现；对于非结构化数据则需借助内容识别引擎来对文件的内容及附属品进行深度认知。其二，建立统一的数据标记体系是实现精确化管控的重要手段。每个数据或一组数据应具备不同种类的标记，如：商业标记、安全标记、科技标记、合法标记等。标记生成可采用自动化标记及人工核对方式：由辨识结果生成的原始标记，然后由数据所有者或者安全专家审核修改。标签体系应该与元数据管理系统、数据目录系统深度集成，实现标签信息随数据流通而流转。

三、大数据环境下敏感数据全生命周期安全管控体系构建

（一）数据采集阶段：源头管控与合规接入

数据采集是敏感信息生命周期管理的第一步，在这一步上做好的工作能够为后续的数据使用奠定良好的基础[7]。因此在数据采集环节应遵循“最小化”的原则，明确不同业务场景下数据采集的内容范围、频率及方式，并杜绝越界采集敏感数据的行为。对于来自外部的数据，应建立合规的接入机制来确认数据来源的有效性和访问范围的有效性以及数据提供者的安全保障能力。在技术实现上，可在采集端口处或数据录入层部署前端识别器，对进入的数据流进行实时处理和敏感性判断，对于超出范围或者过分敏感的数据进行预警或者拦截。针对 UGC 文本、日志等内容应特别关注是否有敏感信息，可加入敏感数据过滤组件在入库前甄别清除；同时建立数据源溯源机制，为每条数据记录附加不可篡改的原始数据印记，用于后期的数据审计及追溯。

（二）数据存储阶段：分级防护与安全加密

对于存储阶段而言，在这个过程中保障数据安全的主要措施是采取与其数据敏感性相适应的安全防护等级：一般数据和公开数据采用低保护级别的访问控制和容错技术；关键数据采用较高保护级别的访问控制技术和完整性技术；重要数据则采取最高保护级别的安全防护技术，如数据加密、设备安全的模块保护以及严格的访问审查 等。对于密码策略的选择，则应根据数据被使用频率及环境而定：若为高频访问的高敏感数据，可选用透明式加密或者列级别加密方式，在保证安全的前提下减少性能损耗；若为低频访问的档案数据，那在推荐上建议采用全盘加密或者文件级加密配合密钥多层级管理机制[8]。对于存储结构来说，可以采用数据分类分区的存储方式，将不同级别的敏感性数据进行物理或逻辑隔离，避免因为混放而引起高级别数据泄漏风险。

（三）数据传输阶段：全程加密与动态管控

数据传输阶段面临窃听、劫持、篡改等安全威胁，其中窃听攻击：攻击者试图获取传输的数据；劫持攻击：攻击者试图将传输过程中的数据截获并进行存储；篡改攻击：攻击者试图对传输的数据进行篡改。对此，管控首要考虑的是如何保护敏感信息不被泄露和破坏，在传输过程中采取最基本的防护手段是端到端加密，即涉及敏感信息的所有通信均采用 TLS/IPsec 加密传输[9]。此外，还应禁止明文传输敏感信息，对于跨域传递不同类型的安全级别信息的情况，比如内外向云、生产到测试等，则应在网络边界部署安全网关进行隧道加密传输以及协议过滤；动态化则希望根据数据等级、传输路径风险、用户行为特征等因素动态地调整安全策略。具体而言，可以构建 SDN 技术下的动态访问控制系统，在检测到存在高风险传输行为后自动执行增强认证、限速或者终止操作；在数据交互场景下如数据共享、公开等场合，应构建安全交换系统，实现审查机制与传输控制的协同作用，此时非法传输请求无法发送。

（四）数据使用阶段：动态脱敏与细粒度管控

在此过程中要做好数据安全和个人信息保护工作，这部分是

整个链条中最敏感也是最容易受到侵害的地方，例如分析、软件测试、业务查询以及共享给第三方等都可能会造成数据泄漏事件的发生。而避免此类事情发生的有效手段之一即为引入动态脱敏技术进行使用期安全管理。该技术的主要思路是在向用户提供数据之前根据用户的身份信息、访问场景及数据敏感性等因素实现对敏感数据的实时转换或隐藏。所以动态脱敏应当具有策略配置功能，在服务器端如客服人员请求客户个人资料时，他们可能只能看到身份证号后四位；但如果是一个数据分析人员想要取出来所有信息的话，那么得到的就是一个彻底打码过的身份证号码了。面对软件测试场景，可以采用静态脱敏的方法，也就是在相对安全的生产环境中进行数据脱敏操作，然后将其拷贝至测试环境中，可以有效防止敏感信息的非保护性披露[9]；二是细粒度访问控制技术能够应对传统基于表格级或者文档级的权限系统难以满足大数据应用场景需求的情况，需要引入基于属性的访问控制系统：并且结合用户特征、数据特征、环境特征及行为特征等多种手段实现行级、列级甚至元素级的访问控制机制[10]。

（五）数据销毁阶段：彻底销毁与合规留痕

擦除意味着信息生命周期的终止，也是最容易被忽视的安全环节。彻底擦除是指对存储介质上的敏感信息进行不可追溯性清除，以避免它们通过任何技术手段恢复。对于机械硬盘，可通过多次覆盖的方式进行信息擦除；而对于 SSD 来说，由于有磨损均衡的存在，简单的擦除无法保证彻底擦除数据，应该使用安全擦除指令或是物理摧毁；对于云计算环境来说，则需要确认所有备份文件、镜像以及缓存都被清除并获取垃圾证明（Cloud Service Provider）。合规的做法就是要留下记录，比如销毁申请、审批过程、执行动作、验证确认等，以备监管审计和司法认可之用；从业务管理上，则应形成分级别的销毁方案，在不同保密等级下有不同的销毁方式及留存时限。

四、结语

大数据环境下敏感数据的识别分类分级与安全管控，是一个涉及技术、管理、组织等多维度的复杂系统工程。本文提出了技术优化的三条路径：构建“规则 +AI”融合的智能识别体系以提升识别的精度和效率，建立“业务 - 安全”双维分类分级框架以实现差异化的风险判定，优化数据资产盘点与标签化管理以夯实治理的数据基础。未来，还需持续推动技术迭代与管理创新，方能在数据要素化浪潮中行稳致远。

参考文献

[1] 杨新涛，曾谛，陈青钦，常津铭，李果 . 数据安全管控体系的研究与应用——基于 " 六步工作法 " 的探索 [J]. 中国宽带 ,2025,21(10):49-51.

[2] 仇建栋 . 智慧校园隐私保护的数据分级策略研究 [J]. 信息系统工程 ,2025,(02):119-123.

[3] 李苹 . 铁路敏感数据流通安全管控体系研究及应用 [J]. 铁路计算机应用 ,2024,33(10):67-72.

[4] 林婧 . 政务数据安全治理研究 [D]. 福建师范大学 ,2024.

[5] 陈易璨 . 侦查中的数据分类分级保护制度研究 [D]. 西南政法大学 ,2023.

[6] 谭柳燕 . 基于 AI 的数据分类关键技术研究 [D]. 四川大学 ,2023.

[7] 安美芳 . 一种面向大数据安全的敏感内容识别机制研究与实现 [D]. 北京交通大学 ,2022.

[8] 张新华 . 大数据安全与敏感数据保护技术应用研究 [J]. 信息记录材料 ,2021,22(04):211-212.

[9] 焦罡 . 大数据环境下敏感数据资产梳理研究 [J]. 中国科技信息 ,2020,(18):76+79.

[10] 焦罡，王飞 . 基于大数据平台的数据安全技术能力体系建设研究 [J]. 信息记录材料 ,2020,21(07):139-140.

