ISSN 1000-0054 清华大学学报（自然科学版） 2024 年 第 64 卷 第 9 期 5/15
CN 11-2223/N J Tsinghua Univ (Sci & Technol), 2024, Vol. 64, No. 9 1565-1576

多源异构空管运行数据安全自动分级模型

陈宝刚[1]， 杨敬轩[1]， 张 毅[1,2]， 晏 松[3]， 何泓霖[1]

（1. 清华大学 自动化系，北京 100084；2. 江苏省现代城市交通技术创新中心，南京 210096；
3. 中国人民公安大学 交通管理学院，北京 100038）

摘 要：随着民航空管信息化建设的持续深化，空管系统作为支撑航空运输业高效、安全运作的关键，面临着数据交换安全性、实时性和高效性的严峻挑战，实现跨行业、跨业务领域、跨组织机构的多源异构空管运行数据安全交换成为当前空管系统信息化建设的重要任务。该文聚焦多源异构空管运行数据安全自动分级模型的构建，旨在通过自动化和智能化的手段对多源异构空管运行数据安全进行自动分级。根据空管运行数据所属业务类型，构造空管运行数据安全分级数据集，并设计13项安全分级特征以全面反映数据安全属性。结合相关法律法规和空管运行数据安全特征，设定5个数据安全级别。基于以决策树为基础分类器的系综算法，建立空管运行数据安全自动分级模型。通过与其他机器学习算法的对比实验，验证了所提算法在自动分级准确率上的优越性，达到了95.5%。

关键词：多源异构数据；空管运行数据；数据安全；自动分级

中图分类号： TP309.2 **文献标志码：** A **文章编号：** 1000-0054(2024)09-1565-10

DOI： 10.16511/j.cnki.qhdxxb.2024.22.036

随着民航空管信息化建设的不断深入，空管系统逐渐成为支撑航空运输业高效、安全运作的关键。大量的业务数据被产生和处理，并通过与外部个体或组织进行跨领域信息交换，满足日益增长的信息化需求。与此同时，数据交换的安全性、实时性和高效性等问题也日益凸显，成为制约空管系统进一步发展的瓶颈。民航空管系统内的应用系统逐渐分立，形成了多个信息孤岛[1]。这不仅阻碍了信息的有效流通，也限制了空管系统整体工作效率的提升。

因此，实现数据安全交换已成为当前空管系统信息化建设的重要任务，关键在于跨行业、跨业务领域、跨组织机构的多源异构空管运行数据安全交换。跨领域信息交换涉及信息存储、元数据注册、用户身份认证、访问控制等多个环节，面临数据多源性、异构性等挑战。

针对跨领域数据安全交换，现有研究已经进行了初步探索。在数据分类分级管理的重要性及实施方法方面，黄伟庆[2]从合规视角出发研究了数据要素的分类分级管理机制，龚钢军等[3]则从网络安全视角审视了配电网数据资产的分类分级。在数据安全治理方面内，罗海宁[4]探讨了在政府数字化转型的背景下建立数据安全堡垒的策略，Tawalbeh等[5]研究了基于数据分级的安全云计算模型，Sabetta等[6]则提出了一种软件系统的数据安全分级方法。此外，数据安全治理在铁路[7]、互联网企业[8]、工业互联网[9]、医学[10]等领域也受到了广泛关注。关于数据安全与开放共享平衡，程军军等[11]提出了在确保数据安全的前提下促进数据开放共享的策略，李玉亮[12]则强调了数据分类分级在数据安全共享中的核心作用。在数据安全治理的法律与制度保障方面，陈家宁等[13]从《数据安全法（草案）》的角度分析了中国跨境数据流动治理方案。在空管数据安全方面，黄长春等[14]针对民航机场协同决策（airport collaborative decision making，A-CDM）系统设计了数据安全保护框架，白宇晨[15]对西北空管局私有云数据平台的网络安全防控体系进行了设计，刘龙庚[16]则对空管安全监控关联算法

收稿日期： 2024-04-24

基金项目： 国家自然科学基金重点项目(62133002)

作者简介： 陈宝刚(1975—)，男，博士研究生。

通信作者： 张毅，教授，E-mail：zhyi@tsinghua.edu.cn

进行研究，设计了空管设备日志关联模型。

如何针对多源异构空管运行数据安全进行有效的自动分级，仍然是亟待解决的问题。空管运行数据安全自动分级的挑战主要体现在数据的多源性和异构性上。空管运行数据涉及多个领域和维度，可能来源于不同的系统和设备，格式和结构也各不相同。另外，空管运行数据的动态性和时效性也给自动分级带来了挑战。空管运行数据中包含实时动态变化的信息，例如动态航迹信息等，这要求自动分级方法能够实时反映数据的变化情况。

本文建立的多源异构空管运行数据安全自动分级模型，通过自动化和智能化的手段，实现了对多源异构的空管运行数据进行安全性自动分级。本文模型依据数据的可用性、完整性和机密性等关键安全属性，对空管运行数据进行自动分级，进而可以为不同安全级别的数据制定相应的安全策略、访问控制机制以及安全管控措施。构建多源异构空管运行数据安全自动分级模型对于提升民航空管系统的数据管理水平、保障数据安全、促进信息共享和业务协同具有重要意义。

在构建多源异构空管运行数据安全自动分级模型的过程中，本文首先构造了一个空管运行数据安全分级数据集。根据空管运行数据所属生产运营业务类型，选取代表性数据，并基于数据安全属性设计了13项安全分级特征。考虑到数据安全分级相关法律法规及民航空管运行数据的特征，本研究定义了5个安全级别。其次，建立空管运行数据安全自动分级模型。基于空管运行数据的特点，以决策树为基础分类器，并采用系综算法，提升自动分级的准确性。通过与其他机器学习算法进行对比，实验结果证明所提模型具有最高的自动分级准确率，达到了95.5%。

1 空管运行数据安全自动分级数据集

为建立多源异构空管运行数据安全自动分级模型，首先需要构造空管运行数据安全自动分级数据集(dataset for automatic classification，DAC)，再利用数据集DAC对可用于数据安全自动分级的机器学习算法进行训练，得到空管运行数据安全自动分级模型。为构造数据集DAC，需要对空管运行数据进行分类与特征分析。具体而言，空管运行数据所属生产运营业务类型及包含的主要信息如下。

1）管制运行：经过扇区信息、经验滑行时间信息、离场信息、离港信息、临时航线信息、流量控制信息、旅客信息、跑道运行模式、实际降落信息、综合航迹信息、4D航迹信息等；

2）通信导航监视：自动化系统航迹信息、管制员通话信息、电报信息、飞行计划信息等；

3）气象：每日天气报告信息、起飞天气预报信息、特殊天气报告信息、重要气象情报信息等；

4）情报：航线数据信息、航段数据信息、航行通告信息、空域数据信息、机场数据信息等；

5）生产运行管理：电子值班网络规划信息、安全策略信息、设备配置信息、日志信息等；

6）综合管理：人员信息、财务信息、固定资产信息、基建信息、办公自动化信息等。

为构造数据集DAC，选择空管运行数据中具有代表性的1 000项数据，占全部数据项总量约10%。具体而言，按照空管运行数据所属六大类业务类型，分别选取10%的代表性数据项来建立数据集DAC。每类业务选取的数据项个数如表1所示。

表1 数据集DAC包含的业务类型及数据量

业务类型名称	数据项个数
管制运行	22
通信导航监视	896
气象	14
情报	35
生产运行管理	29
综合管理	4

1.1 空管运行数据安全自动分级特征

根据空管运行数据的安全属性设计数据集DAC中包含的特征，包括数据可用性A、数据完整性I、数据机密性C等。

1）数据可用性A，是指数据在需要时能够被授权用户或系统访问和使用的能力，涉及数据的存储、备份、恢复和访问控制等方面，对于保障业务的连续性和稳定运行至关重要。

2）数据完整性I，是指数据的准确性和一致性，以及数据未经授权不被修改或破坏，涉及数据的精确性、完整性和一致性等方面。

3）数据机密性C，是指数据不被未经授权的个体或系统访问或泄露，涉及数据的加密、访问控制和隐私保护等方面。为了确保数据机密性，通常需要使用加密算法对数据进行加密，同时实施严格的访问控制策略，限制用户对数据的访问权限。

根据空管运行数据的可用性、完整性和机密性，建立面向空管运行数据安全分级的13项特征，如图1[17]所示。

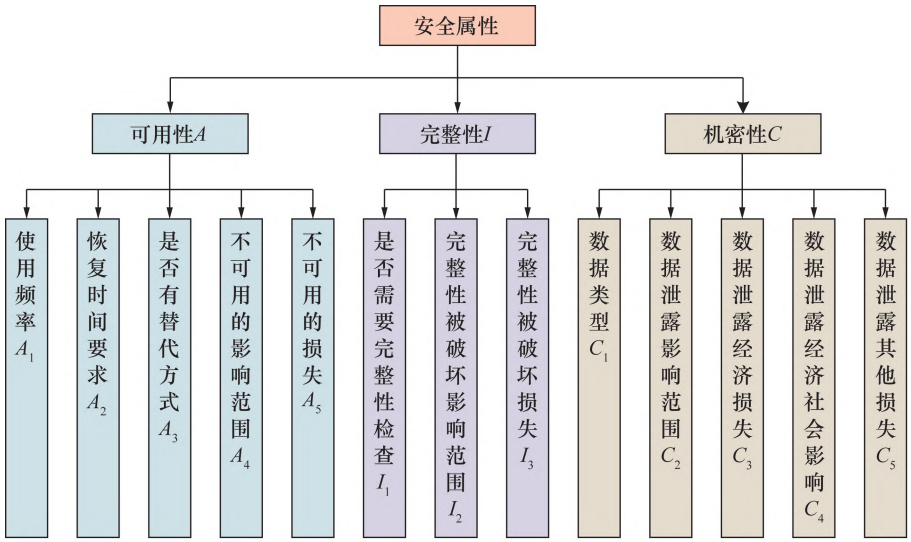

图 1 空管运行数据安全自动分级特征[17]

每个特征的具体含义及可选值[17]如下。

1）使用频率 A_1，反映数据项使用频率高低，可选值为 A_1＝{低/0.25，一般/0.50，频繁/0.75，非常频繁/1.00}；

2）恢复时间要求 A_2，反映数据项出现问题后恢复时间要求，可选值为 A_2＝{长/0.25，一般/0.50，短/0.75，实时切换/1.00}；

3）是否有替代方式 A_3，反映数据项是否有替代数据，可选值为 A_3＝{是/0.00，否/1.00}；

4）不可用的影响范围 A_4，反映数据项不可用的影响范围大小，可选值为 A_4＝{内部/0.50，外部/1.00}；

5）不可用的损失 A_5，反映数据项不可用的损失大小，可选值为 A_5＝{间接/0.25，小/0.50，中/0.75，大/1.00}；

6）是否需要完整性检查 I_1，反映数据项是否需要完整性检查，可选值为 I_1＝{否/0.00，是/1.00}；

7）完整性被破坏影响范围 I_2，反映数据项完整性被破坏影响范围为内部还是外部，可选值为 I_2＝{内部/0.50，外部/1.00}；

8）完整性被破坏损失 I_3，反映数据项完整性被破坏损失大小，可选值为 I_3＝{间接/0.25，小/0.50，中/0.75，大/1.00}；

9）数据类型 C_1，反映数据项的数据类型（公开，内部，敏感，高敏感），可选值为 C_1＝{公开/0.25，内部/0.50，敏感/0.75，高敏感/1.00}；

10）数据泄露影响范围 C_2，反映数据项泄露影响范围为内部还是外部，可选值为 C_2＝{内部/0.50，外部/1.00}；

11）数据泄露经济损失 C_3，反映数据项泄露经济损失大小，可选值为 C_3＝{间接/0.25，小/0.50，中/0.75，大/1.00}；

12）数据泄露经济社会影响 C_4，反映数据项泄露经济社会影响大小，可选值为 C_4＝{间接/0.25，小/0.50，中/0.75，大/1.00}；

13）数据泄露其他损失 C_5，反映数据项泄露其他（非经济）损失大小，可选值为 C_5＝{间接/0.25，小/0.50，中/0.75，大/1.00}。

1.2 空管运行数据安全自动分级标签

基于《数据安全法》《工业数据分类分级指南（试行）》《民航空管系统信息资源管理规定》等相关规定，主要参考《网络安全标准实践指南——数据分类分级指引（征求意见稿）》，并结合空管运行数据安全共享实际情况与需求，根据不同类型的业务与数据共享面临的不同等级的安全风险，建立层次分明的数据安全分级体系，包括 5 个安全级别。

• 1 级，极低共享风险：数据一旦遭到篡改、破坏、泄露或者非法获取、非法利用，可能对个人合法权益、组织合法权益造成轻微危害，但不会危害公共利益、国家安全；

• 2 级，低共享风险：数据一旦遭到篡改、破坏、泄露或者非法获取、非法利用，可能对个人合法权益、组织合法权益造成一般危害，或者对公共利益造成轻微危害，但不会危害国家安全；

• 3 级，中共享风险：数据一旦遭到篡改、破坏、泄露或者非法获取、非法利用，可能对个人合法权益、组织合法权益造成严重危害，或者对公共利益造成一般危害，但不会危害国家安全；

• 4级，高共享风险：数据一旦遭到篡改、破坏、泄露或者非法获取、非法利用，可能对个人合法权益、组织合法权益造成特别严重危害，可能对公共利益造成严重危害，或者对国家安全造成轻微或一般危害；

• 5级，极高共享风险：数据一旦遭到篡改、破坏、泄露或者非法获取、非法利用，可能对公共利益造成特别严重危害，或者对国家安全造成严重或特别严重危害。

1.3 空管运行数据安全自动分级标注

专家标注法是一种基于专家知识和经验的数据标注方法，核心在于利用领域内具有丰富经验和专业知识的专家对特定数据集进行手工标注，从而为后续的模型训练、评估和测试提供高质量的标签数据。考虑到多源异构空管运行数据的专业性，邀请5位专家，采用专家标注法对空管运行数据安全分级数据集的13项特征与安全级别进行标注。以20项数据为例，数据集DAC的标注结果如表2所示。表2中：ADS-B为广播式自动相关监视(automatic dependent surveillance broadcast)，PBCS为基于性能的通信与监视(performance-based communication and surveillance)。

表2 数据集DAC中数据项、特征与安全级别示例

数据项名称	A_1	A_2	A_3	A_4	A_5	I_1	I_2	I_3	C_1	C_2	C_3	C_4	C_5	安全级别
预计起飞机场	1.00.	0.75	1.00	1.00	0.75	1.00	1.00	0.75	0.25	1.00	0.500	0.50	0.50	3
预计降落机场	1.00	0.75	1.00	1.00	0.75	1.00	1.00	0.75	0.25	1.00	0.50	0.50	0.50	3
预计撤轮档时刻	1.00	0.75	1.00	1.00	0.75	1.00	1.00	0.75	0.25	1.00	0.50	0.50	0.50	2
预计飞行时间	0.75	0.75	1.00.	0.50	0.25	1.00	0.50	0.25	0.25	1.00	0.50	0.50	0.50	1
预计航路	1.00	0.75	1.00	1.00	0.75	1.00	1.00	0.75	0.75	1.00	0.75	0.75	0.50	3
预计航空器注册号	0.75	0.75	1.00	1.00	0.50	1.00	1.00	0.50	0.75	1.00	0.75	0.75	0.50	2
预计备降机场	1.00	0.75	1.00	1.00	0.75	1.00	1.00	0.75	0.75	1.00	0.75	0.75	0.50	3
ADS-B能力	0.75	0.75	1.00	0.50	0.50	1.00	0.50	0.50	0.75	1.00	0.75	0.75	0.50	2
PBCS能力	0.50	0.75	1.00	0.50	0.50	1.00	0.50	0.50	0.75	1.00	0.75	0.75	0.50	3
起飞机场	1.00	0.75	1.00	1.00	0.50	1.00	1.00	0.50	0.25	1.00	0.50	0.50	0.50	2
降落机场	1.00	0.75	1.00	1.00	0.50	1.00	1.00	0.50	0.25	1.00	0.50	0.50	0.50	2
实际起飞时间	1.00	0.75	1.00	1.00	0.75	1.00	1.00	0.75	0.25	1.00	0.50	0.50	0.50	2
预计总飞行时间	0.75	0.75	1.00	0.50	0.25	1.00	0.50	0.25	0.25	1.00	0.50	0.50	0.50	1
飞行规则	0.75	0.75	1.00	0.50	0.50	1.00	0.50	0.50	0.75	1.00	0.75	0.75	0.50	3
机载设备与能力	0.50	0.75	1.00	0.50	0.50	1.00	0.50	0.50	0.75	1.00	0.75	0.75	0.50	3
应答机编码	1.00	0.75	1.00	1.00	0.75	1.00	1.00	0.75	0.25	1.00	0.50	0.50	0.50	3
雷达二次编码	1.00	0.75	1.00	1.00	0.75	1.00	1.00	0.75	0.25	1.00	0.50	0.50	0.50	3
预计飞越边境数据	0.75	0.75	1.00	0.50	0.25	1.00	0.50	0.25	0.25	1.00	0.50	0.50	0.50	1
经度	1.00	0.50	1.00	1.00	0.50	1.00	1.00	0.50	0.75	1.00	0.50	0.50	0.50	4
纬度	1.00	0.50	1.00	1.00	0.50	1.00	1.00	0.50	0.75	1.00	0.50	0.50	0.50	4

2 空管运行数据安全自动分级模型

2.1 空管运行数据安全自动分级模型概述

基于机器学习算法构建空管运行数据安全自动分级模型，如图2所示。图2中数据特征 $\boldsymbol{x}=(A_1, A_2, \cdots, C_5)$ 为数据集DAC中的特征项，y^* 为数据的标签，即安全级别。将数据特征 $\boldsymbol{x}$ 输入机器学习算法 f，则该算法可以输出预测的数据安全级别 $y=f(\boldsymbol{x})$。机器学习算法 f 包含可训练参数 w，通过优化机器学习算法预测的安全级别 y 与真实的安全级别 y^* 之间的距离 $d(y, y^*)$ 来迭代更新机器学习算法的参数 w，可以训练得到自动分级模型。

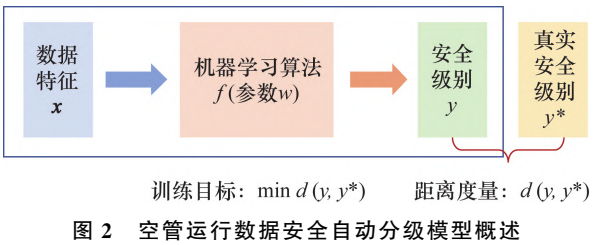

图2 空管运行数据安全自动分级模型概述

2.2 空管运行数据安全自动分级模型

自适应增强算法(adaptive boosting，AdaBoost)

算法[18]是一种系综算法，能够对多个基础分类器进行线性组合，并不断优化组合系数以达到较高的分类准确率。根据空管运行数据安全自动分级数据集 DAC 的特点，选择以分类与回归树(classification and regression tree，CART)[19]为基础分类器的 AdaBoost 算法作为空管运行数据安全自动分级模型。

记输入数据特征和标签为 $T=\{(x_i, y_i)\}_{i=1}^{m}$，其中 x_i 为数据特征(13 项安全属性)，y_i 为数据标签(安全级别)，m 为训练集包含的数据总量。设总迭代次数为 K，并初始化数据集中每项数据的权重为 $\boldsymbol{D}_1=(w_{1,1}, w_{1,2}, \cdots, w_{1,m})$，其中

$$w_{1,i}=\frac{1}{m},\ i=1, 2, \cdots, m.$$

在第 k 次迭代中，$k=1, 2, \cdots, K$，首先使用具有权重 $\boldsymbol{D}_k$ 的数据训练 CART 模型 G_k，则模型 G_k 的分类误差可以表示为

$$e_k=\sum_{i=1}^{m} w_{k,i} I\{G_k(x_i)\neq y_i\}.$$

其中，

$$I\{G_k(x_i)\neq y_i\}=\begin{cases}1, & G_k(x_i)\neq y_i\\ 0, & G_k(x_i)=y_i.\end{cases}$$

设定模型 G_k 的系数为

$$\alpha_k=\frac{1}{2}\ln\frac{1-e_k}{e_k}+\ln(R-1).$$

其中 $R=5$ 为数据标签的类别数量(即 5 个安全级别)。

于是，每项数据的权重可以更新为

$$w_{k+1,i}=\frac{w_{k,i}}{Z_k}\exp[-\alpha_k y_i G_k(x_i)],\ i=1, \cdots, m.$$

其中，

$$Z_k=\sum_{i=1}^{m} w_{k,i}\exp[-\alpha_k y_i G_k(x_i)].$$

最后，得到空管运行数据安全自动分级模型为

$$y=f(\boldsymbol{x})=\mathrm{sign}\left(\sum_{k=1}^{K}\alpha_k G_k(\boldsymbol{x})\right).$$

在对数据进行自动分级时，将待分级数据项的特征 $\boldsymbol{x}'$ 输入训练好的自动分级模型 f，则可获得数据项的安全级别 $y'=f(\boldsymbol{x}')$。

3 实验结果分析

首先，对空管运行数据安全自动分级数据集 DAC 进行统计分析。数据集中安全级别的频数分布如图 3 所示。安全级别为 3 级的数据项最多，而安全级别为 1 级和 5 级的数据项则较少，整体近似正态分布。

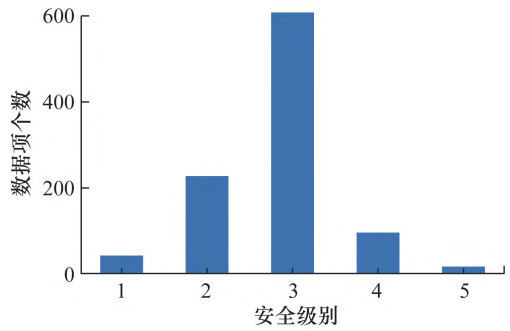

图 3 数据集 DAC 中安全级别的频数分布

将数据集 DAC 按照 4∶1 划分为训练集和测试集，即训练集包含 800 项数据，测试集包含 200 项数据。其中，测试集 200 项数据的安全级别如图 4 所示。

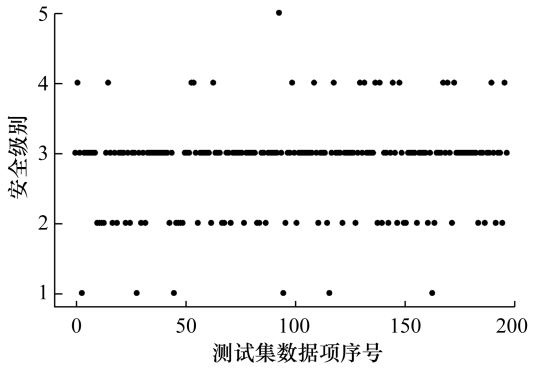

图 4 测试集中数据的真实安全级别

将本文所提方法与 Linear Regression[20]、Ridge[21]、Lasso[22]、Elastic Net[23]、Decision Tree[19]、Random Forest[24]、Extra Tree[25]、Bagging[26]、Gradient Boosting[27]、XGB(extreme gradient boosting)[28]，XGBRF(extreme gradient boosting random forest)[29]等机器学习方法在训练集上进行训练，并在测试集上进行对比。全部 12 种自动分级算法在测试集上的分级结果如图 5 所示。将图 5 与 4 进行对比分析可知，全部算法在测试集中的大部分样本上都预测出了正确的安全级别。

为分析这些算法安全分级错误的情况，绘制全部算法自动分级错误的数据项安全级别(图中显示为彩色圆点)及其真实安全级别(图中显示为灰色圆点)，如图 6 所示。Linear Regression、Ridge、Lasso、Elastic Net 等算法自动分级错误的数据项个数较多，而本文所采用的 AdaBoost 算法的自动分级错误数据项个数最少。

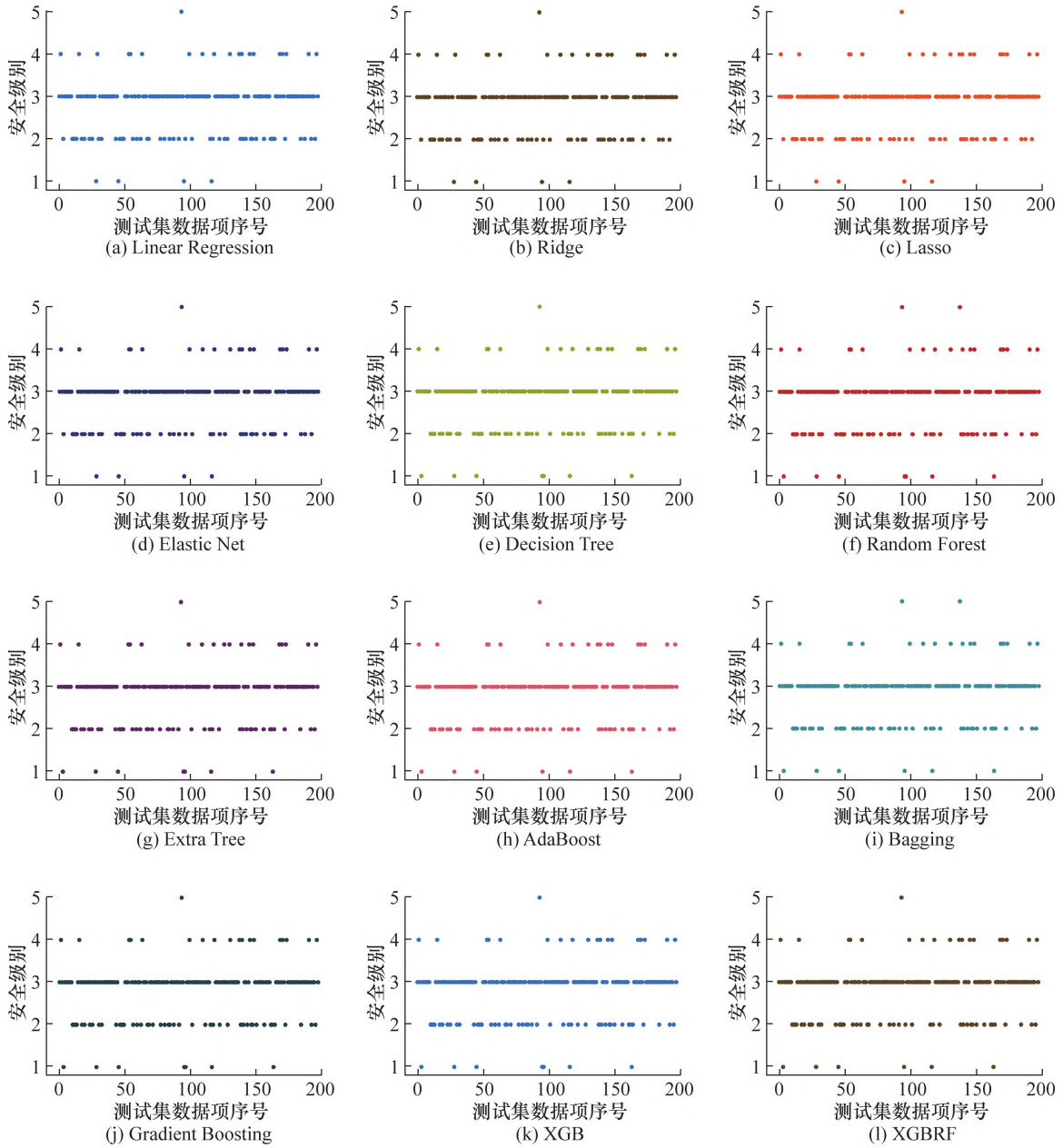

图 5　全部算法在测试集上的自动分级结果

全部算法的自动分级准确率如图 7 所示，其中分级准确率为分类正确的数据项个数与测试集全部数据项个数的比值。Linear Regression、Ridge、Lasso、Elastic Net 等自动分级算法的分级准确率较低，仅有 91%左右，而本文所采用的以 CART 决策树为基础分类器的 AdaBoost 算法的分级准确率最高，可达 95.5%。因此，本文所提针对多源异构空管运行数据安全的自动分级模型可以基于数据安全属性和安全级别，通过在数据集上的训练实现空管运行数据安全的自动分级，且取得了较高的分级准确率。

4　结　论

本文深入研究了在民航空管系统信息化建设的背景下，多源异构空管运行数据的安全交互共享问题，提出并实现了多源异构空管运行数据安全自动分级模型。该模型利用自动化与智能化技术，精准完成了空管运行数据的安全分级，有效提升了空管系统数据的安全管理与交换能力。首先，依据空管运行数据的安全属性和业务特性，构造了一个空管运行数据安全分级数据集，包括 13 个安全分级特征和 5 个安全级别。其次，采用 CART 作为基础分类器，结合 AdaBoost 算法，建立了空管运行数据

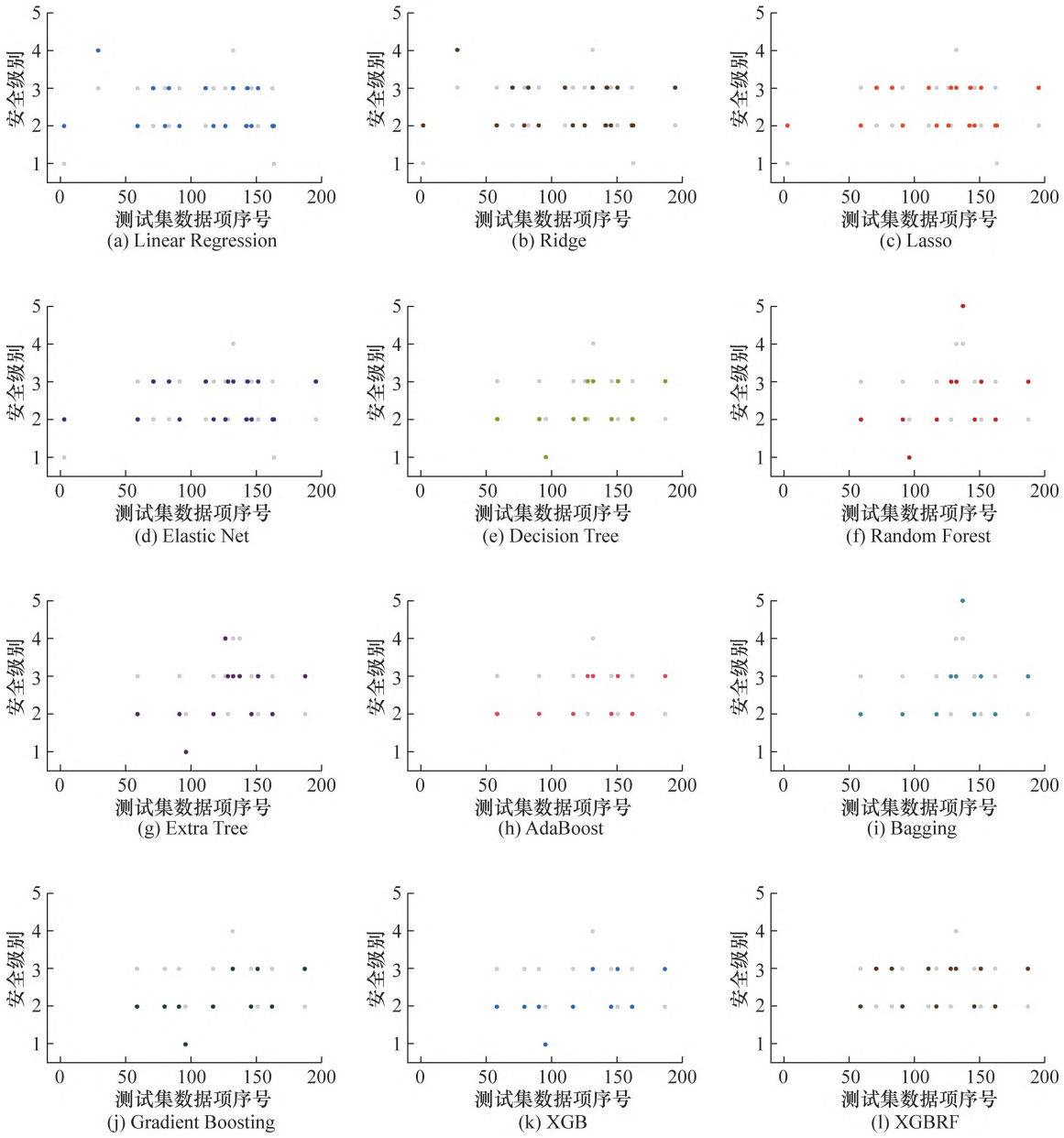

图 6　全部算法自动分级错误的数据项安全级别(彩色圆点)及真实安全级别(灰色圆点)

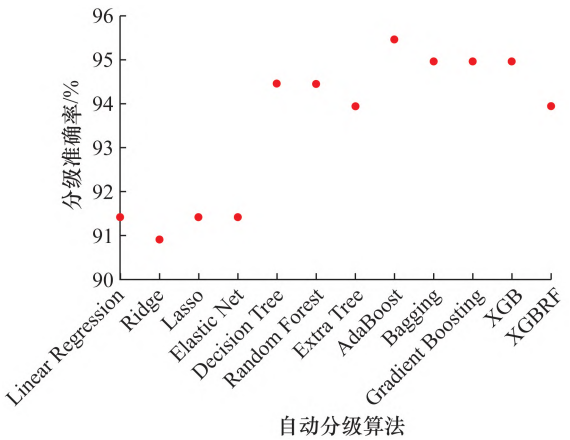

图 7　全部算法的自动分级准确率

的安全自动分级模型。该模型能够根据数据的可用性、完整性和机密性等安全属性，自动且准确地对多源异构的空管运行数据进行分级。最后，通过与其他机器学习算法的对比实验，验证了本研究所提出的模型在自动分级的准确率方面具有优越性，达到了 95.5%的最高准确率。本研究不仅提升了空管数据的安全管理水平，也为民航空管系统的信息安全交换提供了有力支持。

将空管运行数据安全自动分级模型在空管系统中进行实际部署时需要考虑系统的整合性。空管系统是一个非常复杂的系统，在部署自动分级算法时，需要考虑如何与现有系统进行有效整合，避免

对现有系统造成干扰。另外，空管运行数据安全自动分级模型的设计理念和算法框架具有一定的普适性。然而，由于不同行业和领域的数据特性和业务需求存在差异，因此在移植到其他行业领域时需要进行相应的调整和优化。

参考文献 (References)

[1] 戴剑伟，张海粟，王强，等. 跨领域信息交换方法与技术[M]. 2版. 北京：电子工业出版社，2021.
DAI J W，ZHANG H S，WANG Q，et al. Cross domain information exchange methods and technologies [M]. 2nd ed. Beijing：Publishing House of Electronics Industry，2021. (in Chinese)

[2] 黄伟庆. 合规视角下数据要素的分类分级管理机制研究[J]. 上海政法学院学报(法治论丛)，2024，39(2)：121－140.
HUANG W Q. A research on the hierarchical and classified management mechanism of data elements under the compliance perspective [J]. Journal of Shanghai University of Political Science and Law (The Rule of Law Forum)，2024，39(2)：121－140. (in Chinese)

[3] 龚钢军，常卓越，陈志敏，等. 网络安全视角下配电网数据资产分类分级探讨[J/OL]. 华北电力大学学报(自然科学版)，2024：1－14. (2024-03-11) [2024-04-03]. http://kns.cnki.net/kcms/detail/13.1212.tm.20240307.1031.002.html.
GONG G J，CHANG Z Y，CHEN Z M，et al. Discussion on classification and grading of distribution network data assets from the perspective of network security [J/OL]. Journal of North China Electric Power University (Natural Science Edition)，2024：1－14. (2024-03-11) [2024-04-03]. http://kns.cnki.net/kcms/detail/13.1212.tm.20240307.1031.002.html. (in Chinese)

[4] 罗海宁. 政府数字化转型中探索建立数据安全堡垒的实践研究[J]. 中国信息安全，2020(11)：42－45.
LUO H N. Practical research on exploring the establishment of data security fortresses in government digital transformation [J]. China Information Security，2020(11)：42－45. (in Chinese)

[5] TAWALBEH L，DARWAZEH N S，AL-QASSAS R S，et al. A secure cloud computing model based on data classification [J]. Procedia Computer Science，2015，52：1153－1158.

[6] SABETTA A，BEZZI M. A practical approach to the automatic classification of security-relevant commits [C]// 2018 IEEE International Conference on Software Maintenance and Evolution (ICSME). Madrid，Spain：IEEE，2018：579－582.

[7] 彭剑峰，徐保民，张义祥. 基于等保2.0的铁路敏感数据安全关键技术及研究[J]. 网络安全技术与应用，2021(1)：138－142.
PENG J F，XU B M，ZHANG Y X. Key technologies and research on railway sensitive data security based on equal protection 2.0 [J]. Network Security Technology & Application，2021(1)：138－142. (in Chinese)

[8] 胡能鹏，刘晓光. 互联网企业数据安全管理应用研究[J]. 网络安全技术与应用，2020(12)：76－77.
HU N P，LIU X G. Research on the application of Internet enterprise data security management [J]. Network Security Technology & Application，2020(12)：76－77. (in Chinese)

[9] 张雪莹，杨帅锋，王冲华，等. 工业互联网数据安全分类分级防护框架研究[J]. 信息技术与网络安全，2021，40(1)：2－9.
ZHANG X Y，YANG S F，WANG C H，et al. Research on industrial Internet data security classification and grading protection framework [J]. Information Technology and Network Security，2021，40(1)：2－9. (in Chinese)

[10] 刘莉，陈先来，李忠民，等. 精准医学大数据应用安全分类分级研究[J]. 医学信息学杂志，2021，42(1)：9－15，35.
LIU L，CHEN X L，LI Z M，et al. Study on security classification of big data application in precision medicine [J]. Journal of Medical Informatics，2021，42(1)：9－15，35. (in Chinese)

[11] 程军军，杜少雄，姚轶崭. 对数字经济环境下数据安全与开放共享的思考[J]. 中国信息安全，2021(5)：52－54.
CHENG J J，DU S X，YAO Y Z. Reflections on data security and open sharing in the digital economy environment [J]. China Information Security，2021(5)：52－54. (in Chinese)

[12] 李玉亮. 数据分类分级的现状与发展[J]. 中国信息安全，2021(5)：55－56.
LI Y L. Current status and development of data classification and grading [J]. China Information Security，2021(5)：55－56. (in Chinese)

[13] 陈家宁，张建文. 跨境数据流动治理的中国方案：以《数据安全法(草案)》为视角[J]. 长春理工大学学报(社会科学版)，2021，34(2)：35－40.
CHEN J N，ZHANG J W. China's solution for cross-border data flow governance：From the perspective of *Data Security Law* (*Draft*) [J]. Journal of Changchun University of Science and Technology (Social Sciences Edition)，2021，34(2)：35－40. (in Chinese)

[14] 黄长春，齐雅楠. 民航 A-CDM 系统数据安全保护方案探析[J]. 信息安全研究，2023，9(5)：482－489.
HUANG C C，QI Y N. Preliminary study on data security protection scheme of civil aviation A-CDM system [J]. Journal of Information Security Research，2023，9(5)：482－489. (in Chinese)

[15] 白宇晨. 西北空管局基于私有云数据平台的网络安全防控体系设计[J]. 网络安全技术与应用，2022(11)：57－59.
BAI Y C. Design of network security prevention and control system based on private cloud data platform in Northwest Air Traffic Management Bureau [J]. Network Security Technology & Application，2022(11)：57－59. (in Chinese)

[16] 刘龙庚. 基于异构网络空管安全监控关联算法研究[J]. 信息网络安全，2022，22(4)：58－66.
LIU L G. Research on association algorithm of heterogeneous network security monitoring [J]. Netinfo Security，2022，22(4)：58－66. (in Chinese)

[17] 黄洪，刘增良，余达太. 一种智能化的数据分类、分级及保护模型[J]. 北京工业大学学报，2011，37(6)：921－927.
HUANG H，LIU Z L，YU D T. An intelligent model of data classification and protection [J]. Journal of Beijing University of Technology，2011，37(6)：921－927. (in Chinese)

[18] ZHU J，ROSSET S，ZOU H，et al. Multi-class AdaBoost [J]. Statistics and Its Interface，2009，2(3)：349－360.

[19] LOH W Y. Classification and regression trees [J]. WIREs Data Mining and Knowledge Discovery，2011，1(1)：14－23.

[20] MONTGOMERY D C，PECK E A，VINING G G. Introduction to linear regression analysis [M]. 6th ed. Hoboken，USA：John Wiley & Sons，2021.

[21] MCDONALD G C. Ridge regression [J]. WIREs Computational Statistics，2009，1(1)：93－100.

[22] RANSTAM J，COOK J A. LASSO regression [J]. British Journal of Surgery，2018，105(10)：1348－1348.

[23] ZOU H，HASTIE T. Regularization and variable selection via the elastic net [J]. Journal of the Royal Statistical Society Series B：Statistical Methodology，2005，67(2)：301－320.

[24] RIGATTI S J. Random forest [J]. Journal of Insurance Medicine，2017，47(1)：31－39.

[25] SHARAFF A，GUPTA H. Extra-tree classifier with metaheuristics approach for email classification [M]// BHATIA S，TIWARI S，MISHRA K，et al. Advances in computer communication and computational sciences：Proceedings of IC4S 2018. Singapore：Springer，2019：189－197.

[26] BREIMAN L. Bagging predictors [J]. Machine Learning，1996，24(2)：123－140.

[27] NATEKIN A，KNOLL A. Gradient boosting machines，a tutorial [J]. Frontiers in Neurorobotics，2013，7：21.

[28] CHEN T，HE T，BENESTY M，et al. XGBoost：Extreme gradient boosting [Z]. R Package Version 0.4－2. 2015：1－4.

[29] ZHANG W G，WU C Z，ZHONG H Y，et al. Prediction of undrained shear strength using extreme gradient boosting and random forest based on Bayesian optimization [J]. Geoscience Frontiers，2021，12(1)：469－477.

Automatic classification model for multisource heterogeneous air traffic control operational data security

CHEN Baogang[1]，YANG Jingxuan[1]，ZHANG Yi[1,2]，YAN Song[3]，HE Honglin[1]

(1. Department of Automation，Tsinghua University，Beijing 100084，China；
2. Jiangsu Province Collaborative Innovation Center of Modern Urban Traffic Technologies，Nanjing 210096，China；
3. School of Traffic Management，People's Public Security University of China，Beijing 100038，China)

Abstract：[Objective] With the continuous advancement of the informationization of air traffic control (ATC) in civil aviation，the ATC system currently acts as a hub supporting the efficient and safe operation of the aviation transportation industry. In this process，a large volume of business data is generated and processed within the ATC system that needs to be exchanged across different domains with external entities or organizations to meet the growing demands of informatization. However，data security，real-time processing，and efficiency issues have become increasingly prominent，posing bottlenecks to the further development of the ATC system. Driven by the promotion of informationization of the ATC system，the application subsystems within the civil aviation ATC system have gradually become fragmented，forming multiple information silos. This

not only hinders the effective circulation of information but also limits the overall operational efficiency of the ATC system. Therefore, facilitating information sharing and system integration has become a critical task in the current phase of informationization. The exchange of information across industries, business domains, and organizations is a key aspect of achieving these goals. The process of cross-domain information exchange is considerably more complex than simply transmitting information from one place to another, involving multiple stages such as information storage, metadata registration, user identity authentication, and access control. Moreover, cross-domain information exchange also faces many challenges, including data heterogeneity, platform heterogeneity, distribution, autonomy, and security. This study aims to address these challenges by proposing a model for the automatic classification of multisource heterogeneous ATC operational data security to enhance data management, ensure security, promote information sharing, and facilitate business collaboration within the civil aviation ATC system. [Methods] Herein, first, a dataset is constructed to facilitate the classification of the ATC operational data security. Representative data from various operational categories are selected, and 13 key security attributes are identified to design the data security classification. Five security levels are established based on relevant laws and regulations pertaining to data security and the characteristics of the civil aviation ATC operational data. Subsequently, an automatic classification model is developed based on the AdaBoost algorithm with the classification and regression tree (CART) as the base classifier, considering the unique characteristics of the ATC operational data. [Results] Experimental results demonstrate the effectiveness of the proposed automatic classification model. A comparative analysis of the proposed model against other machine learning algorithms reveals that the proposed model achieves the highest accuracy rate, reaching 95.5%. Thus, the proposed model successfully classifies multisource heterogeneous ATC operational data according to their security attributes, enabling the formulation of tailored security strategies and access control mechanisms for different data security levels. [Conclusions] This proposed model considerably enhances the data management capabilities of the civil aviation ATC system, ensures data security, promotes information sharing, and facilitates business collaboration within the system. Thus, this study provides a robust framework for addressing the challenges associated with data security and integration in complex operational environments, laying a foundation for further advancements in civil aviation ATC informationization.

Key words: multisource heterogeneous data; air traffic control operational data; data security; automatic classification

(责任编辑 李丽)