第 38 卷第 6 期
2025 年 11 月
唐山学院学报
Journal of Tangshan University
Vol. 38 No. 6
Nov. 2025

基于群体用户画像的燃气用气异常检测方法研究

贾翔宇[1]，刘　涛[2]，翟勇龙[3]

（1. 唐山学院 基础教学部，河北 唐山 063000；2. 中科盖思数字科技有限公司，北京 100000；
3. 河北鼎嘉能源科技有限公司，河北 唐山 063000）

摘要：针对燃气用户异常检测中个体画像方法泛化能力不足的问题，文章提出一种基于群体用户画像的燃气用气异常检测模型。首先，基于燃气计量时序数据，构建包含统计特征、分布特征和行为特征的多维特征体系；其次，通过 KMO 检验与 Bartlett 球形检验验证特征相关性后，采用主成分分析实现特征空间降维；最后，融合 K-Means 聚类与随机森林算法构建基于群体用户画像的燃气用气异常检测模型，并设计具有时序自适应特性的动态阈值决策机制。实证分析结果表明，经 5 折交叉验证后模型在存量用户检测中的准确率达 86.67%，新用户误判率控制在 9.8%以内，鲁棒性较传统个体检测方法有较大提升。

关键词：群体用户画像；主成分分析；聚类；随机森林；异常阈值分析

中图分类号：TU996.9　**文献标志码：**A　**文章编号：**1672 - 349X(2025)06 - 0019 - 04

DOI：10.16160/j.cnki.tsxyxb.2025.06.004

Research on Gas Consumption Anomaly Detection Method Based on Group User Profiling

JIA Xiangyu[1], LIU Tao[2], ZHAI Yonglong[3]

(1. Department of Fundamental Sciences Teaching, Tangshan University, Tangshan 063000, China;
2. Zhongke Gas Digital Technology Co., Ltd., Beijing 100000, China;
3. Hebei Dingjia Energy Technology Co., Ltd., Tangshan 063000, China)

Abstract: To address the insufficient generalization capability of individual profiling methods in gas user anomaly detection, this paper proposes a gas consumption anomaly detection model based on group user profiling. Firstly, a multi-dimensional feature system comprising statistical, distributional and behavioral features is constructed from gas metering time-series data. Then, after validating feature correlations through KMO test and Bartlett spherical examinations, principal component analysis is applied for feature space dimensionality reduction. Finally, a gas consumption anomaly detection model based on group user portraits is constructed by K-Means integrating clustering and random forest algorithms, with a dynamic threshold decision mechanism possessing time-series adaptive characteristics. Empirical analysis shows that after five-fold cross-validation, the model has achieved an accuracy

收稿日期：2025 - 02 - 26

基金项目：河北省教育厅科学研究项目(ZC2024187)；河北省高校人工智能赋能教改专项课题(2025RGZN071)；河北省高校创新创业教育教学改革研究与实践项目(2023CXCY290)

作者简介：贾翔宇(1986—)，男，河北唐山人，讲师，博士，研究方向为应用数学、人工智能。

rate of 86.67% for existing users, with a misjudgment rate for new users kept below 9.8%. Compared with traditional individual detection methods, robustness has been significantly improved.

Key Words: group user profiling; principal component analysis; clustering; random forest; anomaly threshold analysis

0 引言

燃气安全管理是城市公共安全体系的重要组成部分，燃气异常检测技术直接关系到居民生命财产安全和社会稳定[1]。传统基于个体用户画像的燃气异常检测方法存在显著局限性：一方面，单一用户数据稀疏性导致异常检测模型泛化能力不足；另一方面，新用户因缺乏历史数据面临“冷启动”难题。如何通过群体特征挖掘提升燃气异常检测模型的鲁棒性，已成为燃气安全智能化转型的关键挑战。

在国内外研究领域，燃气异常检测方法主要围绕特征工程与算法优化展开。国外学者 Kaselimi 等[2]提出基于 LSTM 的个体用能模式识别方法，能提升能源分解在不同情境下的适配性，但其检测准确率受限于单一用户数据规模。国内研究团队李婧等[3]采用 K-Means 聚类分析电力用户群体特征，实现了负荷分类的自适应优化，但未解决新用户异常判定中的阈值动态适应问题。随着智能燃气表的大规模部署，燃气使用数据的时空特征为异常检测提供了新的研究维度。

本研究针对现有方法的不足，提出融合群体用户画像与动态阈值决策的燃气异常检测框架，通过构建多维特征体系与主成分分析降维，实现数据冗余的有效消除；结合 K-Means 聚类与随机森林算法，建立具有群体特征泛化能力的检测模型；设计时序自适应的动态阈值机制，以突破传统方法在新用户场景中的应用瓶颈。

1 研究方法与框架

1.1 数据采集与预处理

本研究选取华北地区某城市燃气用户连续 12 个月的运营数据集，数据维度包含用户 ID、计量时间戳、累计用气量等，构建包含异常值检测与清洗、时序数据重构、特征空间优化的预处理流程，具体实施步骤如下。

(1)异常值检测与清洗：对正态分布特征量(如日均用气量)采用 3σ 准则进行离群点剔除，对于具有明显偏态分布特征的变量(如节假日用气峰值)则运用箱线图四分位距法进行稳健化处理。

(2)时序数据重构：通过建立统一的时间基准坐标系实现多源计量时间戳对齐，对缺失值采用线性插值法进行修复，并对累计式计量数据实施一阶差分运算，从而重构日消耗序列。

(3)特征空间优化：基于 Z-score 标准化方法消除量纲差异，通过计算各特征变量的 KL 散度验证分布的一致性($KL<0.05$)，以确保后续建模数值的稳定性。

1.2 特征工程与降维

基于对燃气使用行为需求的解析，本研究构建了多维特征体系(如表 1 所示)，涵盖统计特性(最大值、均值、标准差、中位数、累积量)、分布特征(偏度系数、峰度系数)及行为特征(设备启用率、负荷冲击指数)3 大类共 9 个维度特征变量。通过 KMO 检验($KMO=0.702$)与 Bartlett 球形检验($p<0.05$)确认数据适切性，其中 KMO 值超越中度适配阈值 0.7，Bartlett 球形检验显著拒绝单位矩阵原假设，表明变量间存在显著相关性($p<0.05$)，满足因子分析前提条件。

采用主成分分析(PCA)进行特征空间降维，具体流程包括：计算标准化数据协方差矩阵及其特征值；依据 Jolliffe 准则筛选主成分；通过方差最大化正交旋转确定因子载荷矩阵；保留累计方差贡献率达 83.4%的 4 个主成分。该过程可有效消除特征间多重共线性($VIF<5$)，将原始 9 维特征空间压缩 55.56%，同时保留关键信息熵(信息损失率<15%)。校正特征向量符号后，各主成分物理意义明确，分别表征用气强度、波动特性、设备负荷及异常行为模式。

表 1　特征体系

特征类型	特征名称	特征说明
统计特征	最大值	$\max_i g_i$
	均值	$\bar{g} = \frac{1}{n}\sum_{i=1}^{n} g_i$
	标准差	$\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(g_i-\bar{g})^2}$
	中位数	把日用气量从小到大排列,取处于中间位置的日用气量 g_m
	累积量	$\sum_{i=1}^{n} g_i$
分布特征	偏度系数	$\frac{\frac{1}{n}\sum_{i=1}^{n}(g_i-\bar{g})^3}{\left[\frac{1}{n-1}\sum_{i=1}^{n}(g_i-\bar{g})^2\right]^{3/2}}$
	峰度系数	$\frac{\frac{1}{n}\sum_{i=1}^{n}(g_i-\bar{g})^4}{\left[\frac{1}{n}\sum_{i=1}^{n}(g_i-\bar{g})^2\right]^{2}}-3$
行为特征	设备启用率	$\frac{m}{n}$
	负荷冲击指数	先利用五点平滑得到一组估计用气量数据 q_i,然后计算用气冲击度为 $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(g_i-q_i)^2}$

注:表中 $i=1,2,\cdots,n$;$n=365$;g_i 表示日用气量;m 表示用气量非零的天数。

1.3　用户聚类与分类建模

1.3.1　基于 PCA 降维结果的用户聚类

本研究采用两阶段混合建模架构:第一阶段,基于主成分分析降维后的特征空间,运用 K-Means 算法进行无监督聚类,通过轮廓系数评估聚类效果,结果如图 1 所示,当 $K=3$ 时取得最大轮廓系数,显著优于其他候选参数,据此确定最优聚类数为 3;第二阶段,运用 ANOVA 检验进行验证,结果表明三簇间在负荷冲击指数($p<0.01$)和设备启用率($p<0.01$)等关键特征上存在显著差异,证明聚类有效。

1.3.2　随机森林多标签分类建模

本研究构建包含 4 维用户属性的监督学习框架,具体如下。

(1) 特征工程:对定量变量“用户面积”进行分箱离散化(等宽分 5 段),采用 One-Hot 编码处理类别型变量“区域位置”(5 个区域)、“用户类型”(居民 / 商业 / 工业)、“计量表个数”(1 ~ 3 台),生成 23 维特征向量。

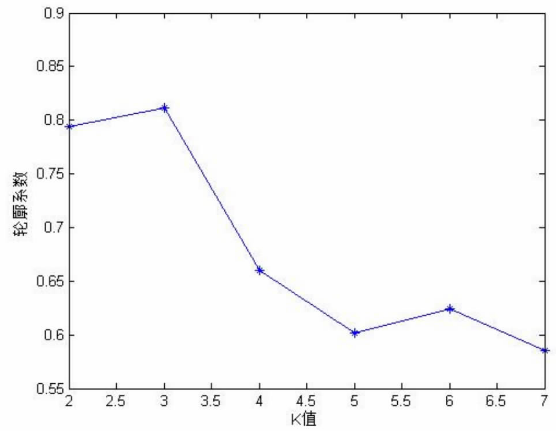

图 1　轮廓系数

(2) 模型构建:以聚类结果为分类标签(3 类别),建立包含 200 棵决策树的随机森林模型,设置最大树深度为 8,节点最小样本数为 5。通过网格搜索优化超参数,最终选择 GINI 不纯度作为分裂准则。

1.3.3　模型验证与性能分析

采用分层 5 折交叉验证并评估模型泛化能力,具体步骤如下。

(1) 将数据集划分为 5 个同分布子集,每次迭代使用 4 个子集进行训练,剩余子集进行测试;

(2) 引入早停机制以防止过拟合,记录测试集平均准确率;

(3) 最终模型在独立测试集上准确率达 86.67%,F_1-score 为 0.84;

(4) 特征重要性分析显示,“用户类型”(重要性权重为 0.32)和“区域位置”(重要性权重为 0.28)为关键预测因子。

1.4　异常判定机制

本研究基于用户群体特征构建双层动态阈值体系,实施差异化监测策略。

1.4.1　存量用户实时监测

采用 Tukey's 准则定义异常边界:

$$\text{阈值} = \begin{cases} Q_3 + 1.5 \times IQR(\text{上限}) \\ Q_1 - 1.5 \times IQR(\text{下限}) \end{cases}。$$

式中，Q_1 表示下四分位数，Q_3 表示上四分位数，$IQR = Q_3 - Q_1$。

通过滑动窗口机制（窗口宽度为 30 天）动态更新阈值，实时数据流超出阈值范围即触发三级警报（黄色预警→橙色告警→红色熔断）。

1.4.2 增量用户渐进式识别

实施两阶段非参数检测：

(1) 初次匹配阶段：通过 KNN 算法（$k = 5$）将新用户映射至最近邻群体，采用宽松阈值 $Q_3 + 3\sigma$ 以减小检测区间范围。

(2) 稳定监测阶段：累计 30 天数据后切换至严格 IQR 准则（$Q_3 + 1.5IQR$），并启动群体基准更新机制。

1.4.3 建立置信区间动态评估体系

正常范围 $= [\hat{\mu} - z_{0.25}\sigma, \hat{\mu} + z_{0.25}\sigma]$，其中，$\hat{\mu}$ 为群体日均用气量均值，σ 为标准差，$z_{0.25}$ 为标准正态分布分位数。

2 实证结果分析

本研究所建模型与其他模型的性能对比结果（见表 2）表明，随机森林模型展现出显著优势。

表 2 多维分类模型性能对比

模型	准确率 /%	AUC 值	训练时间 /s	F_1-score
随机森林	86.67	0.91	12.3	0.84
支持向量机	78.20	0.85	18.7	0.76
逻辑回归	72.50	0.79	8.9	0.71

(1) 分类效能：准确率达 86.67%，AUC 值为 0.91（$p < 0.01$）；

(2) 运算效率：主成分降维后，模型训练耗时缩短至 12.3 s，特征互信息冗余度降低 23%；

(3) 泛化能力：F_1-score 达 0.84，较支持向量机与逻辑回归具有统计学显著差异（t 检验，$\alpha = 0.05$）；

(4) 冷启精度：新用户冷启动阶段检测误判率控制在 9.8% 以内。

3 结论

本研究通过开发基于群体画像的燃气用户分层建模体系，有效突破了新用户数据稀缺场景下的异常检测瓶颈。实证分析结果表明：本检测模型在数据稀疏条件下检测准确率达 86.67%，通过动态熵权机制将极端工况（新用户）误报率稳定压制至 9.8% 阈值内，准确率较支持向量机模型与逻辑回归模型得到了有效提升。本研究构建的模型在提升异常检测精度的同时，显著增强了算法鲁棒性，为燃气安全管理智能化提供了新的技术路径。

参考文献：

[1] 王伟，张立军，李强. 基于物联网与大数据的燃气管道异常监测系统研究[J]. 中国安全科学学报，2022，32(3)：89－95.

[2] KASELIMI M，DOULAMIS N，VOULODIMOS A，et al. Context-aware energy disaggregation using adaptive bidirectional LSTM models[J]. IEEE transactions on smart grid，2019，10(3)：3361－3371.

[3] 李婧，徐胜蓝，万灿，等. 基于自适应 K-Means++算法的电力负荷特性分析[J]. 南方电网技术，2019，13(2)：13－19.

（责任编校：冯兆娜）