

DOI: 10.13955/j.yzyj.2022.01.07.05

大数据在逾限邮件智能分类分级中的应用

冯涛, 王辉, 杨超丰, 邓爱丽, 毕道玉, 李丹凤

(中国邮政集团有限公司河南省分公司, 河南 郑州 450000)

摘 要: 介绍了河南邮政寄递业务邮件时限管理中的“标杆管理”理论, 以寄递网提质增效为中心, 立足省情实际, 坚持问题导向, 精准科学施策, 用大数据技术推动寄递环节时限达标创标, 提升寄递业务网运质量和运营能力。

关键词: 决策树; 逾限邮件; 时效; 提质增效

中图分类号: F61 **文献标识码:** A

为积极推动寄递业务飞速发展, 促进寄递网转型升级和邮件运行质量提高, 河南邮政根据中国邮政集团有限公司提出的“三个视角一个提升”的指示精神, 引入“标杆管理”理论, 将邮件运行时限与同业领先者对标, 以逾限邮件为抓手, 经过科学建模、细化算法、深挖逾限环节、洞察逾限原因, 以大数据视角为寄递业务抓重点、补短板、强弱项、固优势的目标提供数据支撑。

1 时限管理现状

1.1 对标竞品差距明显

与竞品相比, 河南邮政部分线路在运行时限方面存在一定差距, 客户满意度不高。通过对标准快递和快递包裹主要线路运行时限分析发现, 邮件在省内时限整体落后于竞品。按环节分析, 主要问题集中在处理中心环节, 因处理能力不足导致邮件整体时限受到影响; 其次是投递环节, 多次投递

以及错分误发等原因导致邮件整体时限受到影响。

1.2 人工分析效率不高

传统的邮件时限运行质量管控主要以人工分析为主, 采用不定时随机抽查的方式。这种方式存在以下不足: 一是效率低、耗时长, 分析结果具有一定的滞后性, 往往错过最佳管控期; 二是分析不全面, 主要依赖经验, 受主观认知制约, 具有一定局限性; 三是共性问题总结不到位, 片面性较大, 易将零散、碎片化的问题认定为主要因素, 无法对运行时限实施针对性的整改提供科学指导。

1.3 缺乏有效管控抓手

河南邮政共设 113 个县(市)分公司, 邮路 1 600 余条, 其中市县邮路 229 条, 县以下投递段道近 6 000 条。因邮件流程复杂、影响因素多, 运营管理工作难度大, 亟需通过大数据手段为寄递时限把脉问诊, 深入剖析每一环节运营质量不佳的根源, 找痛点、抓整改, 以点带面实施靶向提速。

获奖情况: 2020 年全国邮政企业科技创新成果二等奖。

作者简介: 冯涛(1977~), 男, 河南新乡人, 硕士, 高级经济师, 主要从事邮政市场协同发展研究; 王辉(1975~), 男, 河南周口人, 硕士, 主要从事信息化建设及成果推广研究; 杨超丰(1981~), 男, 山西运城人, 经济师, 主要从事大数据分析研究; 邓爱丽(1980~), 女, 甘肃天水人, 高级工程师, 主要从事大数据建模研究; 毕道玉(1984~), 男, 河南商丘人, 高级工程师, 主要从事大数据算法应用研究; 李丹凤(1984~), 女, 河南周口人, 经济师, 主要从事寄递流程优化与时限运行管控研究。

收稿日期: 2021-11-03

本刊网址: yzyj.sjzpc.edu.cn

2 分析准备

2.1 邮件分类

邮件根据流向分为省际邮件和省内互寄邮件。省内互寄邮件除了常规邮件外, 还包含同城、政务、电商邮件, 对不同类别邮件设定相应的分析规则。

同城邮件是指收寄和投递为同一地市的邮件; 电商邮件是指直属、东区、南区、西区、小米、跨境六个机构发寄的邮件, 该类邮件协议客户集中, 时效性强, 对菜鸟指数影响较大, 需要重点分析; 政务邮件是指在营业网点办理的警邮、税邮等便民服务类邮件, 该类邮件的办理和收寄大部分由两个机构完成, 需要单独分析。

2.2 邮区划分

根据业务实际情况, 将河南省 18 个地市分公司划分为 5 个市级邮区, 市级邮区内互寄邮件无需经转省级处理中心, 市级邮区间邮件可直发或经转省级处理中心, 邮区划分见图 1。

根据邮件网络运营标准, 当邮件量达到一定标准后, 可设直发邮路, 目前省内内有 20 余条动态直发邮路。部分县市由于区位原因需单独调整, 如舞钢市隶属于平顶山市, 因其毗邻漯河, 辖区邮件须经转漯河。

2.3 邮件流程划分

流程划分主要围绕邮件省内段开展研究分析, 按照邮件流向分为出口段(上行)和进口段(下行)两部分, 详见图 2。

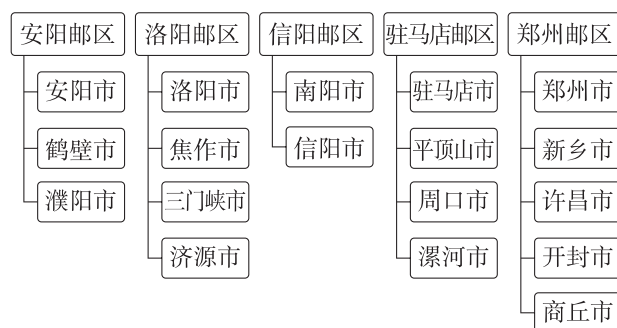


图 1 邮区划分图

如图 2 所示, 邮件上下行共分为 9 个环节, 对每个环节进行剖析, 根据网络运营标准梳理出各环节作业频次和处理时长, 建立运营标准库。

3 模型构建

3.1 决策树

采用经典决策树分类算法对邮件各环节的异常情况进行判别, 主要步骤如下。

3.1.1 样本提取

分类的准确率很大程度上由训练集决定, 训练集的质量对预测结果有较大影响, 为保证训练集充分可用, 在数据库中利用开窗函数 `row_number() over()` 对数据按环节进行分区排序, 按照抽取样本量 3 倍初步选取数据, 利用 `order by dbms_random.value` 和 `sample(m)seed(n)` 随机抽样 10 000 条数据, 其中 70% 为训练集, 30% 为验证集。

3.1.2 特征指标选取和预处理

样本数据主要包含处理时长、在途时长和运行时刻三大类, 利用 R 语言 `cor()` 函数做相关性分

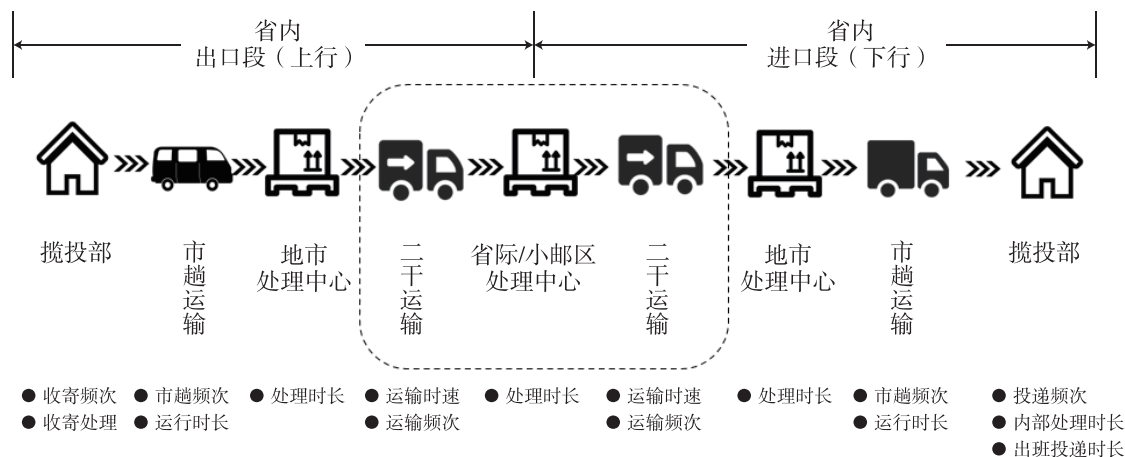


图 2 邮件流程图

析, 衡量变量之间的依赖性强弱, 排除强相关指标, 最终确定下段时长、处理中心处理时长、封发时刻等 16 个特征指标, 对指标进行数值化和标准化处理, 采用中位数对缺失值进行填充, 同时剔除异常值。

3.1.3 属性选择

属性选择方法主要有信息增益、增益比率和基尼指数, 为让每个分支记录的类别纯度更高, 选择信息增益方法。指标属性确定后, 采用信息增益算法确定最好的分裂根节点, 信息增益基于香农的信息论, 具体算法为: 假设选择在途时长作为分裂属性, 训练集中在途时长有 K 个不同的取值 $\{V_1, V_2, V_3, \dots, V_k\}$, 计算按在途时长分裂后的信息量:

$$\text{Info}(D) = -\sum_{i=1}^m p_i \times \log_2(p_i)$$

利用 R 语言 ID3 算法, 参数 `parms=list(split,prior,loss)`, 其中 `split="information"`, 按照纯度最大原则得出内部处理时长为最佳分裂属性, 可作为决策树的根节点。

3.1.4 剪枝调优

生成决策树过程中可能会出现“过拟合”现象, 需要通过剪枝处理提高泛化能力, 剪枝的过程即为自下而上遍历决策树, 寻找子节点均为叶节点的节点, 然后计算剪枝前后惩罚函数是否变小, 若变小则剪枝, 将该节点的输出设为样本中同种最多的类别, 本项目主要采用复杂度参数对决策树进行剪枝调优。

图 3 为复杂度参数 (CP), 用于惩罚过大的树, 虚线是基于一个标准差准则得到的上限为 $0.537\ 5+0.037\ 2$, 从图像看, 应选择虚线下方最左

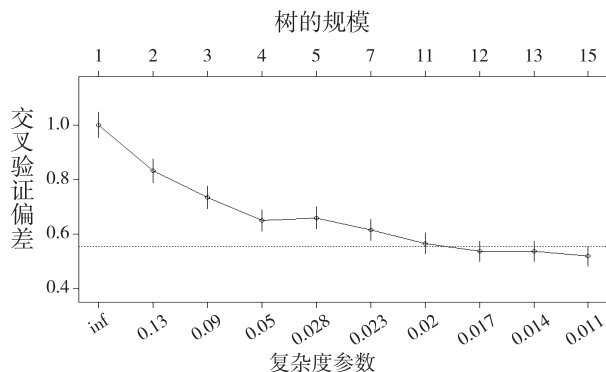


图3 复杂度参数图

侧 CP (0.018 750) 值对应的树。

3.1.5 交叉验证

利用 `Table()` 函数对训练集和验证集进行交叉验证, 准确率达到 94%, 值得注意的是, 基于某个时段样本的决策树模型可能不稳定, 因为在实际业务中的微小变化可能会导致不同的树生成, 不能覆盖所有问题, 需要结合业务相关规范进行调整和优化, 最终形成了分类结果集, 作为本次分析的结论性成果。

3.2 抽样数据决策树分类结果集

根据抽样数据生成的决策树模型见图 4。从树的顶端开始, 以内部处理时长作为根节点, 如果条件成立, 则从左枝往下分裂, 否则从右枝往下, 当观测点到达终端节点时, 分类结束。

利用 R 语言 `asRules` 获取更详细的决策树节点规则, 部分结果集如下:

Rule number: 13 [Kyphosis=absent cover=412 (4%) prob=0.62]

```
INNER_T1<19
DLV_705C<0.5
FWFW_DESC<3.5
INNER_T1<8
WAY_TIME>=9.5
```

Rule number: 25 [Kyphosis=absent cover=209 (2%) prob=1.00]

```
INNER_T1<19
DLV_705C<0.5
FWFW_DESC>=3.5
TO_PKP_T>=46
CURRENT_T>6
```

Rule number: N

以规则 13 为例, 主要描述县市及城郊 (FWFW_DESC<3.5) 揽投范围内未妥投邮件, 反馈距下段间隔不足 5 小时 (DLV_705C<0.5), 在县域处理中心处理时长 (INNER_T1) 在 8 小时以内, 市趟 (WAY_TIME) 超过 9.5 小时, 则运输环节存在异常, 该规则样本覆盖率为 4%, 准确率为 62%, 规则基本符合业务逻辑, 同时结合寄递业务规范和运营标准库, 将结果整理成收寄、上行、经转、下

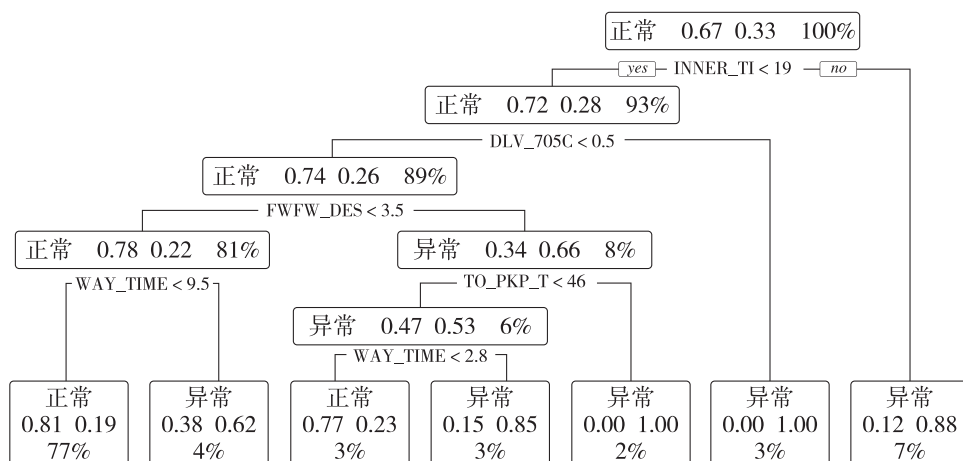


图4 规则集决策树

行、投递五大部分，主要规则如表1所示。

3.3 时限回溯评价分析

以邮件轨迹信息表 JDP_JDPT_QPS_TRACE 中的数据为依据，按月为单位，对机构至机构间运行时长进行分析，对比运行计划，寻找差距，评估邮路运行效率，检验邮路设置科学性和项目运行质量（见表2）。

4 应用效果

该项目自落地应用以来，通过完善优化算法模型不断提高分析能力，日均分析1.54万件逾限邮件，实现了省际省内逾限邮件全覆盖，节约了人力成本，在管理决策和服务支撑两方面取得了良好的应用效果，短板线路时限提升明显，省内网络组织更加合理，时限管理更加精细化。

4.1 管控措施逐步细化

根据项目应用结论，运营管理部指挥调度中心加强时限管控措施：一是业务部门坚持定期通报制度，坚持日通报、周分析、旬推进、月总结，每天对问题环节情况进行通报；二是强化业务流程督导，深入落后单位剖析原因，协助抓好生产流程优化，及时解决发展中遇到的问题与困难，切实提高邮件运行效能；三是坚持典型示范带动，营造比学赶超氛围，安排先进单位分享经验，及时推广先进单位的最佳实践，鼓舞先进、鞭策后进。

4.2 落地效果突出

项目自应用以来，按照“邮路一条一条地捋、环节一环一节地掰、要素一个一个地剖”的要求，以点带面，稳扎稳打。一是提升管理措施，信阳市

区投递局的服务范围与南湾管理区投递部、六里棚投递部因服务范围有重叠，致使邮件错分误发频率较高，通过录入邮件白名单方式，逐步减少邮件错分情况。二是动态调整邮路，项目在中部区域提速上效果显著，新增调整中部直达一干邮路28条，其中新增4条，调整运行时速14条，调整运行时刻16条。新增信阳—长沙（湘临1）、洛阳—武汉（豫临1）等线路，线路开通后信阳邮区至长沙隔日递率提升23.02个百分点，洛阳邮区至武汉提升22.83个百分点，商丘至武汉隔日递率提升23.63个百分点。

4.3 综合效益明显

项目自应用以来，在运营管理部门督导落实下，寄递时限方面取得了良好成效：一是省际重点区域时限水平稳定可控，中部区域特快专递菜鸟达成率87.78%，区域排名第二，快递包裹菜鸟达成率64.47%，区域持续排名第一；二是省内互寄次日递率和及时妥投率提升明显，特快次日递率提升25.22个百分点，快包次日递率提升45.5个百分点，特快及时妥投率达96.95%，同比提升13.08个百分点，快包及时妥投率达98.27%，同比提升6.11个百分点。

5 推广前景展望

该项目基于经典决策树分类算法，对逾限邮件异常环节进行分类分级，通过一系列精准的管控措施，为寄递时限管理工作带来了阶段性成果，在利用大数据支撑业务发展方面积累了经验，对决策管理和网运组织起到重要支撑作用，后续结合寄递

表 1 决策树分类结果表

环节	处理时长	市趟运输时长	发运计划	问题件
收寄	当天收寄的邮件，当天未封发	/	/	
上行	进出口不在同一天 地市封发时间距收寄时间超 H1 小时	超计划时长， 郑州市趟超 X1 小时		
经转	郑州出口邮件，省处理中心时长超 H2 小时 郑州进口邮件，省处理中心时长超 H3 小时 郑州投递邮件，早 7:00 前未封发 市级邮区邮件，在经转机构超过 H4 小时	超过计划时长	标快或快包错装、混装； 晚点超 M 分钟以上	撤单 退件 错分 改址 迁址 待收费
下行	超过 H5 小时（积压延时） 0:00 ~ 3:00 进局的邮件 7:00 前未发车 3:00 ~ 11:00 进局的邮件 14:00 前未发车	超过计划时长		
投递	当天解车，当天未下段 当天下段，当天未反馈 多次投递 妥投后撤回	郑州市趟超 X2 小时 其余市趟超 X3 小时	/	

表 2 机构间运行时长与计划时长差异表

封发机构	封发机构名称	接收机构	接收机构名称	实际时长（小时）	邮件量（万件）	与计划差（小时）
4500**	汽车邮件转运站	4550**	安阳包件处理班	4.79	5.40	0.50
4500**	汽车邮件转运站	4570**	濮阳市邮件处理中心	4.09	4.47	0.03
4500**	汽车邮件转运站	4580**	鹤壁处理中心	4.06	1.45	-0.40
4500**	汽车邮件转运站	4610**	许昌包件处理班	3.44	4.14	0.30
4500**	汽车邮件转运站	4620**	漯河市邮件处理中心	3.22	2.71	0.20
4500**	汽车邮件转运站	4630**	驻马店市转运站	4.62	7.54	0.32

业务发展状况，通过优化算法模型，实现时限管控工作的事前预测、事中纠偏、事后评估，从根本上解决时限不稳定问题。同时，拓展大数据技术应用场景，特别是在业务经营决策方面进行适配和探索。

5.1 深化行业对标，构建市场优势

时限是快递行业最基本的竞争要素，是获取客户的“敲门砖”和“生命线”，通过大数据分析手段，建立竞品时限库，通过线路和环节对标分析，提高弱势线路质量，巩固和扩大优势线路范围，进而向经营端传导，将时限优势转化为市场竞争力，强化源头获客，构建市场优势。

5.2 围绕流程优化，实现降本增效

该算法不仅适用于时限管理分析，还适用于处理中心作业流程优化和选址、运输调度管理、投递段道撤并等应用场景。未来将重点聚焦处理、运

输、投递三大应用场景，通过优化串行运输、邮路套跑、单改双等方式促进运输组织改革，通过提升直投、转窗投、自提网络能力建设方式促进揽投网改革，挖掘压降节点，构建成本优势，提升寄递网运营管控能力。

参 考 文 献

[1] 夏慧玲，潘骁，罗清. 大数据在快递行业中的应用探析 [J]. 江苏商论，2019（6）
[2] 刘志勇，齐立云，原景成. 大数据时代我国快递业的发展现状与趋势 [J]. 交通运输研究，2019（5）
[3] 李敏杰. 基于大数据下的寄递物流管理信息系统的研究 [D]. 南京：南京邮电大学，2014
[4] 毛国君. 数据挖掘原理与算法 [M]. 北京：清华大学出版社，2005