基于机器学习的烟草水分智能控制数据预处理研究

史成云　崔文波　崔汝念　邹欣延

（红塔烟草（集团）有限责任公司昭通卷烟厂，云南 昭通 657000）

摘　要　在实际生产过程中，我们得到的原始数据往往非常混乱、不全面，机器学习模型往往无法从中有效识别并提取信息。数据和特征决定了机器学习的上限，而模型和算法只是逼近这个上限而已，在采集完数据后，机器学习建模的首要步骤以及主要步骤便是数据预处理。本文以基于机器学习的烟草水分智能控制为例，研究其数据的采集、清洗和特征化，以避免因“脏数据”影响机器学习模型预测烟草水分的有效性、可重复性和泛化能力，从而影响模型的质量。

关键词　机器学习　烟草水分　数据采集　数据清洗　特征工程

中图分类号：TP31　　**文献标识码**：A　　**文章编号**：1007-0745(2022)12-0028-03

1 数据采集和清洗

1.1 数据采集

通过 PID 控制器和传感器采集制丝生产线各工序中各生产要素数据。

1.2 数据清洗

1.2.1 删除工序无关特征

根据采集到的数据，分析实际工艺流程中实际影响烟草水分控制的各因素。初步删除对水分控制影响不大的部分特征，保留“设备状态”“设定加水流量”“设定热风温度”“A 组设定加料流量”“B 组设定加料流量”“循环风温度设定”“A 组实际加料流量”“B 组实际加料流量”“实际循环风温度”“实际加水流量”“实际热风温度”“加水累计量”“直喷蒸汽累计量”“直喷蒸汽实际流量”“直喷蒸汽设定流量”“批运行”“批次号”“配方号”“设定出口水分”“实际出口水分”“实际入口水分”“出口测温仪温度”“电子秤累计量”“电子秤实际流量”“电子秤设定流量”“A 模块加料实际流量”“B 模块加料实际流量”“C 模块加料实际流量”“入口温度”“入口湿度”“出口温度”“出口湿度”“时间”“当前生产模块号”共 34 个特征。

1.2.2 删除意义不明确，不具备解释性的特征

进一步观察数据，发现“批运行”特征只有 1 个值，无意义；“作业号”取值为整数，表示当天的生产轮次，且存在 19.29% 的缺失值。删除这两个特征。

“A 组设定加料流量”“B 组设定加料流量”“A 组实际加料流量”“B 组实际加料流量”“A 模块实际加料流量”“B 模块实际加料流量”“C 模块实际加料流量”传感器距离润叶加料桶太远，时滞不好计算，且加料过程是根据烟叶物料量、工艺要求和传送带速度等因素决定的一个均匀的添加过程，所以这些因素对烟叶生丝水分控制不产生直接影响，删除这些特征。

“设定热风温度”“循环风温度设定”“电子秤设定流量”的取值只有 1 个，无分析价值。

“直喷蒸汽设定流量”虽然针对不同配方号的烟叶设定值不同，但对同一配方的烟叶在生产过程中设定值基本一致，且直接作用于烟叶的实际数值应该是“直喷蒸汽实际流量”，所以删除“直喷蒸汽设定流量”特征，同理删除“设定加水流量”特征。

1.2.3 删除空采样数据

每个批次的生产开始前，机器开机预备，控制器和传感器此时亦会采集数据，此时的数据称为空采样数据。

空采样：生产线未投入实际生产时，监测设备进行采样称为空采样，此时获取的数据称为空采样数据。

这些数据对后续分析会产生不必要的影响，删除。

首先按照“配方号、批次号”将每一个生产轮次的数据分隔开，按照每一个生产轮次数据中“电子秤累计量”第 1 个大于 0 的数据作为标志，之前的数据看作是空采样数据。

1.2.4 特征变化和添加

根据烟叶制丝生产的工艺原理，和传感器实际作用，对部分会影响到烟叶生丝含水率并可以推算的特征例如“电子秤瞬时量”“加水瞬时量”“蒸汽瞬时

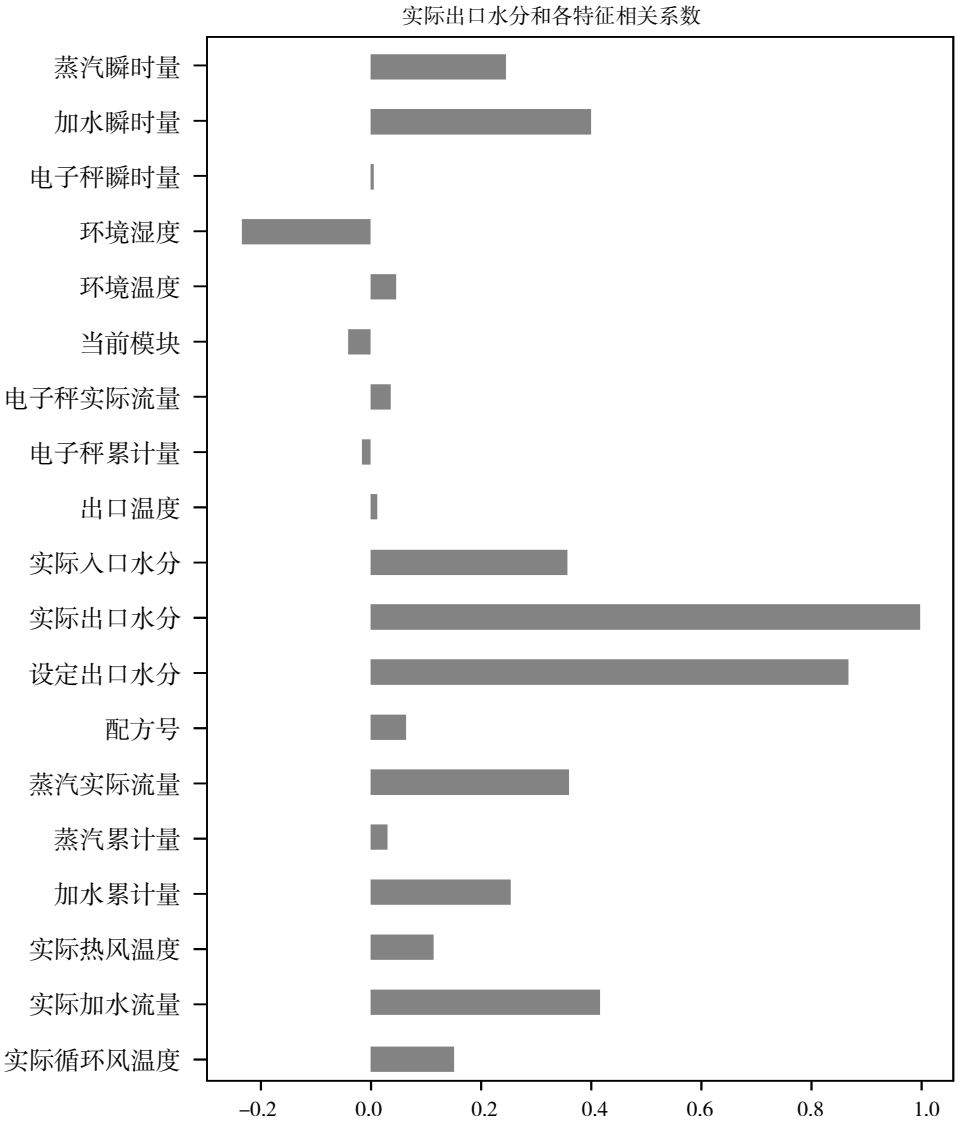

图 1

量”特征，“入口温度”“入口湿度”“出口湿度”“出口温度”分别取其平均作为环境温湿度数据。

对按照“批次号”和“配方号”分割后的数据分别进行计算。

为方便后续表达，对特征名称进行简化：

“实际加水流量”：“实际加水流量”。

“实际热风温度”：“实际热风温度”。

“加水累计量”：“加水累计量”。

“直喷蒸汽累计量”：“蒸汽累计量”。

“直喷蒸汽实际流量”：“蒸汽实际流量”。

“电子秤累计量”：“电子秤累计量”。

“电子秤实际流量”：“电子秤实际流量”。

“出口测温仪温度”：“出口温度”。

“当前生产模块号”：“当前模块”。

1.2.5 时滞数据对齐

在回潮工序共产生了两处时滞：一是烟叶物料从电子秤经过到进入滚筒开始喷水的时滞，此处的时滞通过判断每个批次生产数据中实际加水流量和加水累计量均大于 0 的第一个数据行作为标志，计算时滞 [1]；二是烟叶物料入滚筒后到出滚筒的时滞，从数据观察很难找到统一的标志信息，根据现场多次人工实测定义该时滞 [2]。

1.2.6 异常数据查找和删除

设定条件：（1）整个生产批次的实际出口水分最大值小于 10[3]；（2）整个批次中采样数据时间间隔超过 2 秒；（3）整个生产批次均没有实际加水流量和加

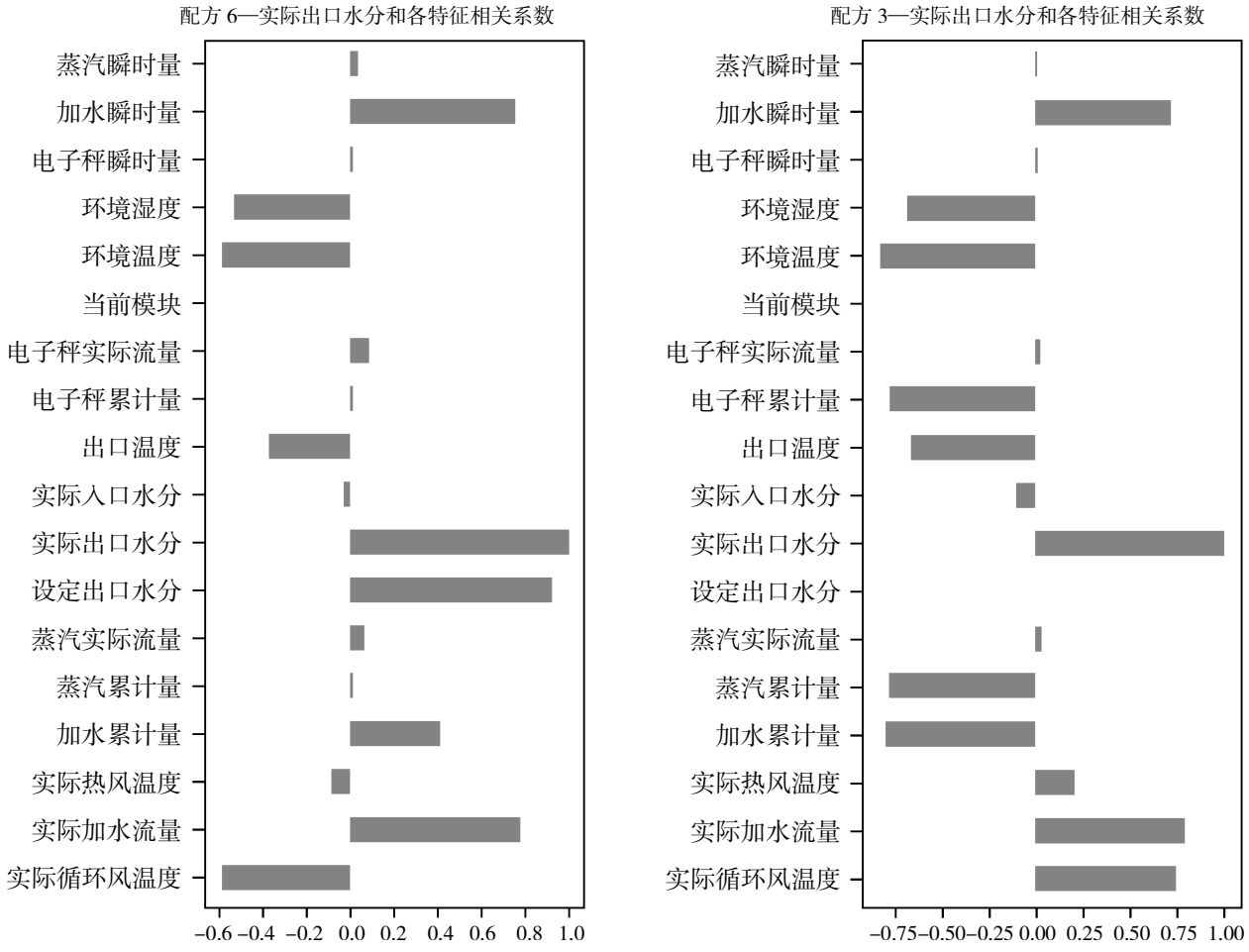

图 2

水累计量均大于 0 的记录 [4]。

符合以上三个条件中的任何一个批次的数据都属于异常数据，删除数据。

1.2.7 去除料头料尾数据

料头和料尾因为物料量不稳定，对应传感器检测值有较大误差，会影响到数据分析。根据生产工艺相关要求前 120 公斤物料看作是料头，时移对齐后出现空值的行看作是料尾，加水瞬时量或电子秤瞬时量为 0 的数据也看作是料尾。

2 特征工程

将上述清洗后的数据，重新拼接成一个大的数据集后考察特征之间的相关性、共线性等因素进行特征选择 [5]。

从总体数据来看，各特征和实际出口水分的相关系数如图 1。

按照不同配方分别分析不同特征和实际出口水分之间的相关性，可以观察到不同配方的烟叶物料在生产过程中，环境温湿度和实际出口水分的相关性有明显变化。

参考文献：

[1] 顾亮，等 . 环境温湿度对配送烟丝质量及卷烟质量的影响 [J]. 食品与机械 ,2017,33(04):190−194.

[2] 国家烟草专卖局 . 卷烟工艺规范 [M]. 北京 : 中央文献出版社 ,2003.

[3] 张云飞，袁鹏，董云，等 . 环境温湿度对制丝水分控制的影响——基于红河卷烟厂制丝过程数据 [J]. 统计学与应用 ,2015(02):34−46.

[4] 陈良元 . 卷烟生产工艺技术 [M]. 郑州 : 河南技术出版社 ,2002.

[5] 卷烟工艺与设备编写组 . 卷烟工艺与设备 [M]. 北京 : 轻工业出版社 ,1988.