ICS 35.240.70
CCS B 07

DB61111

杨凌农业高新技术产业示范区地方标准

DB 6111/T 197—2023

智慧农业园区数据处理技术规范

Technical Specifications for Data Processing of Intelligent Agriculture Parks

2023－04－27 发布　　2023－05－27 实施

杨凌示范区市场监督管理局　　发 布

前　言

本文件按照GB/T 1.1—2020《标准化工作导则　第1部分：标准化文件的结构和起草规则》的规定起草。

本文件由西北农林科技大学信息工程学院提出。

本文件由杨凌示范区农业标准化技术委员会归口。

本文件起草单位：西北农林科技大学（信息工程学院、信息化管理处）、陕西省农村科技开发中心、杨凌耘尚田园网络科技有限公司、杨凌乾泰电子科技有限责任公司、杨凌现代农业产业标准化研究推广服务中心。

本文件主要起草人：刘　斌、耿　楠、蒲　攀、周兆永、张宏鸣、李书琴、黄铝文、刘运松、耿耀君、李　梅、张海曦、卫　星、邓希廉、李皓、马军妮、文立红。

本文件首次发布。

智慧农业园区数据处理技术规范

1 范围

本文件规定了基于物联网系统的智慧农业园数据处理的相关术语和定义、处理流程及技术要求。

本文件适用于智慧农业园的数据规范处理、数据分析、数据可视化管理。

2 规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中，注日期的引用文件，仅该日期对应的版本适用于本文件；不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

GB/T 36344-2018 信息技术 数据质量评价指标

GB/T 37025-2018 信息安全技术 物联网数据传输安全技术要求

QX/T 628-2021 常规高空气象观测数据处理方法

3 术语和定义

下列术语和定义适用于本文件。

3.1

农业大数据 big data of agriculture

在农业生产、经营、管理、服务等环节中产生的海量的，具备多样性、高增长率、真实性及一定价值的数据集。

3.2

数据处理 data processing

为从大量杂乱无章、难于直接理解的数据中抽取并推导到对特定人群有价值、有意义的信息，而进行的数据采集、存储、检索、加工、变换和传输等活动。

3.3

数据增强 data augmentation

一种为提升数据利用价值，用有限数据创造出尽可能多有用信息的数据扩充技术。

3.4

监测预警 monitoring and early warning

通过监测、汇集、抽取相关信息数据，结合风险评估分析，将可能出现的结果直观展现给决策者，提醒作出预先处置的活动。

3.5

特征衍生 feature derivatives

利用现有的数据特征，组合构建出新的数据特征的技术手段。也称特征构建。

3.6

特征选择 feature selection

按系统特定指标最优化目标，从原始特征中选择出最有效特征的过程。也称特征子集选择或属性选择(Attribute selection)。

4 处理流程

4.1 数据预处理

4.1.1 数据清洗

4.1.1.1 检验重复性数据需要根据具体场景、数据特点和问题需求，确定重复性数据的定义、检验方法以及工具，并记录检验结果。

4.1.1.2 异常值采用删除、替换、离群值检测等方法处理。其中离群值检测可以采用箱线图、3σ准则、聚类等统计方法处理。

4.1.1.3 重复值采用删除、合并、标记等方法处理。其中合并可以采用相似度匹配、聚类等方法处理。

4.1.1.4 不一致值采用规范化、转换、匹配等方法处理。其中规范化可以采用大小写转换、数据类型转换等方法处理。

4.1.1.5 格式不一致采用规范化、转换、格式化等方法处理。其中规范化可以采用格式化字符串、正则表达式等方法处理。

4.1.1.6 噪声数据处理包括但不限于以下方法：

a) 采用分箱法、聚类法、回归法等处理噪声数据。

b) 采用分箱法将原始数据划分为若干区间，统计每个区间内的样本数量并计算样本占比，然后平滑处理每个区间的样本占比，将平滑后的数据作为处理后的结果，用于后续的分析和建模。

c) 采用聚类法根据一定的相似性度量分组原始数据，计算每个组的中心点，并根据中心点重新分配数据点到各个组中，迭代执行上述步骤直到满足停止条件为止。最终得到的聚类结果可以用于分析和建模，去除噪声数据对后续分析和建模的影响。

d) 采用回归法包括建立一个回归模型，通过拟合已知数据的函数关系，预测未知数据，并评估和调整预测结果，得到更准确的预测结果。在建立回归模型时，需要去除或修正噪声数据，提高模型的预测精度。

4.1.1.7 清洗处理的数据再次传输应符合 GB/T 37025-2018 的安全传输规定。

4.1.1.8 高空气象数据的处理应符合 QX/T 628-2021 的规定。

4.1.2 数据补全

a) 均值/中位数/众数填充：对于数值型数据，可以使用均值、中位数或众数来填充缺失值。

b) 固定值填充：对于某些特殊的数据，可以使用固定值来填充缺失值。

c) 向前/向后填充：对于时间序列数据，可以使用向前或向后的值来填充缺失值。

d) 插值法填充：可以使用插值法来填充缺失值，例如线性插值、多项式插值等。

e) 建模预测填充：可以使用其他变量建立模型，预测缺失值。

4.1.3 数据融合

采用深度学习模型提取多个数据源的数据信息特征（园区小气候数据、种植环境数据、图像数据和农技知识数据），融合特征级、决策级，提升数据的有效性和准确性。

4.1.4 数据变换

4.1.4.1 图像数据按照目标程度划分为一般或严重状态，通过数字图像处理技术生成充足的数据集，按照 3:1:1 的比例划分为训练集、验证集和测试集。

4.1.4.2 园区小气候数据和种植环境数据，采用特征衍生和特征选择作预处理。特征选择分别基于最大信息系数的最小冗余、最大相关指标过滤粗筛特征，再基于嵌入法选择。

4.1.5 数据规约

应在保证数据集完整性的基础上，简化数据集，包含但不限于以下方面：

a) 维度规约：即将数据的维度降低，减少数据中不必要的属性，降低数据集的复杂度。常用的维度规约方法有主成分分析（PCA）和线性判别分析（LDA）。
b) 数值规约：即将数据的数值范围缩小，减少数值间的差异，降低数据集的复杂度。常用的数值规约方法有归一化和标准化。
c) 属性规约：即从数据集中选择出最具有代表性和区分性的属性，剔除无关或冗余的属性，减少数据集的复杂度。常用的属性规约方法有逐步回归法和决策树算法。
d) 数据压缩：即通过压缩算法去除数据集中的冗余信息，减少数据集的复杂度。常用的数据压缩方法有哈夫曼编码和 Lempel-Ziv 算法。
e) 随机抽样：即从数据集中随机选择部分数据作为样本，通过分析处理，减少数据集的复杂度。常用的随机抽样方法有简单随机抽样和分层抽样。

4.2 数据分析

应建立在计算机机器学习基础上。被分析的数据质量按照GB/T 36344-2018的规定评价，评价合格的数据分析结果为有效分析结果。分析技术包括但不限于：

a) 描述性统计：通过对数据的描述性统计分析，了解数据的基本情况，包括中心趋势、离散程度、分布形态等等。
b) 探索性数据分析：通过可视化和统计分析探索数据之间的关系趋势，为后续建模和分析做准备。
c) 假设检验：用于检验数据之间是否有显著差异，判断是否拒绝某个假设。
d) 回归分析：用于分析自变量和因变量之间的关系，建立回归模型，预测因变量的变化趋势。
e) 聚类分析：将数据分成若干类别，同一类别内的数据相似度较高，不同类别之间的数据相似度较低。
f) 主成分分析：将多个变量合并成少数几个新变量，保留原始变量的大部分信息，以减少维度。
g) 时间序列分析：用于分析时间序列数据，找出趋势、季节性和周期性等规律，预测未来的变化趋势。
h) 关联规则挖掘：从数据中发现频繁出现的模式和关联关系，如超市商品的购买关系。
i) 决策树分析：基于样本数据建立决策树模型，通过选择属性值，逐步筛选出目标属性预测值。
j) 神经网络分析：用人工神经网络模拟人脑处理信息的过程，进行数据建模和预测等操作。

4.3 数据可视化

数据分析与预测结果应以图像、图表的直观方式，展示数据所蕴含的信息、规律与趋势，并可实现交互式处理。

a) 明确对象。通过数据的来源、属性，明确可视化的具体对象。
b) 可视化映射。应选择直观的、易于理解的方式，将数据蕴含的信息呈现给用户。宜使用的数据图表包括但不局限于饼图、柱形图、折线图、条形图等，图表应包含必要的说明注释。
c) 用户感知。可视化后的数据，应具备通过与可视模块间的交互，实现主动获取信息的功能。