doi:10.6041/j.issn.1000-1298.2022.01.009

基于多条件时间序列的免耕播种机作业数据清洗方法

姜含露[1,2]　周利明[3]　马　明[4]　李　阳[3]　周　燕[1,2]　苑严伟[3]

(1.中国科学院半导体研究所,北京 100083;2.中国科学院大学材料科学与光电技术学院,北京 100049;
3.中国农业机械化科学研究院土壤植物机器系统技术国家重点实验室,北京 100083;
4.吉林省农业机械化管理中心,长春 130062)

摘要:为提高作业监测数据状态预测精度,并保证无效数据的实时清洗,提高数据质量并降低监测设备的缓存压力,从而降低对后续地块作业质量评价准确性的影响,减轻数据并发带来的网络压力,本文针对免耕播种机长时序的田间周期性作业规律,提出基于多条件时间序列分析的监测数据清洗方法及模型,该模型包含 3 个长短时记忆特征提取模块,分别提取了工况参数中车速、瞬时面积和播种量的时空特征,再利用通道融合(CONCAT 连接)保证了融合后的特征具有个体差异性。通过该模型可以实时判断当前时刻的免耕播种机工况时序状态值,实现了某位置点作业工况的状态预测,从而间接判断图像抓拍系统的实时清洗状态。40 次迭代后不同模型的对比结果表明:多条件特征通道融合的时间序列模型对有效点和无效点的预测精度都超过了 85%,抓拍图像清洗平均准确率为 92.4%。因此,本文的研究方法以免耕播种机工况状态作为抓拍图像清洗依据是有效的,数据清洗后约有 63% 的冗余数据被剔除。

关键词:免耕播种机;监测;数据清洗;状态预测;时间序列;工况参数

中图分类号:S24　　**文献标识码**:A　　**文章编号**:1000-1298(2022)01-0085-07　　**OSID**:

Data Cleaning Method of No-tillage Seeder Monitoring Data Based on Multi-conditional Time Series

JIANG Hanlu[1,2]　ZHOU Liming[3]　MA Ming[4]　LI Yang[3]　ZHOU Yan[1,2]　YUAN Yanwei[3]

(1. *Institute of Semiconductors*, *Chinese Academy of Sciences*, *Beijing* 100083, *China*
2. *College of Materials Science and Opto-Electronics Technology*, *University of Chinese Academy of Sciences*, *Beijing* 100049, *China*
3. *State Key Laboratory of Soil – Plant – Machine System Technology*,
Chinese Academy of Agricultural Mechanization Sciences, *Beijing* 100083, *China*
4. *Agricultural Mechanization Management Center of Jilin Province*, *Changchun* 130062, *China*)

Abstract: Improving the prediction accuracy of working state of no-tillage seeder and cleaning the invalid data timely will improve the data quality and reduce the cache pressure of monitoring equipment. However, as the agricultural machinery moved back and forth in the farmland, monitoring equipment captured a large number of invalid images at both ends of the farmland or when the vehicle stopped. These images affected the accuracy of farmland operation quality evaluation and created congestion in transmission network. A data cleaning method based on multi-condition time series, mainly vehicle speed, seeding rate and instantaneous area, was proposed to deal with the periodic change of long time series of agricultural machinery in the farmland. The model included multiple long-short term memory (LSTM) and spatiotemporal feature channel fusion (CONCAT connect) to maintain the individual difference under multi-condition. The current time sequence state of the agricultural machinery working condition can be predicted, and the real-time cleaning state of the image capture system can be indirectly acquired. Due to screen and capture valid image from captured image every three minutes by cleaning state, the system achieved the maximum efficiency in transmission channel and memory space. The comparison results of different models after 40 iterations showed that the prediction accuracy of this

收稿日期:2021-01-08　修回日期:2021-02-14

基金项目:国家重点研发计划项目(2016YFD0700305)和兵团重大科技项目(2018AA00404)

作者简介:姜含露(1991—),女,博士生,主要从事农业信息化技术研究,E-mail:halo91@163.com

通信作者:周燕(1972—),女,研究员,博士生导师,主要从事弱光探测与成像研究,E-mail:zhouyan@semi.ac.cn

method for both valid and invalid samples was over 85% and the average accuracy of image cleaning was 92.4%. The data cleaning results showed that about 63% of the redundant data was removed after data cleaning. Therefore, the research method took the working condition of no-tillage seeder as the basis of image cleaning was effective, which had high research value and application prospect.

Key words: no-tillage seeder; monitor; data cleaning; state prediction; time series; working parameters

0 引言

在我国北方地区,秸秆还田免耕播种是一项普遍应用的培肥地力的保护性耕作技术,在免耕播种机上安装监测设备,可以借助高清摄像头抓拍田间秸秆覆盖情况,并进行秸秆覆盖率定量计算。2020 年,吉林省农业机械化远程精准作业平台中保护性耕作的监测面积约为 1.2×10^6 hm^2[1],获取了大约 7.2×10^6 幅抓拍图像。但由于播种机在农田地头间做往返运动,两端调头或长时间停车占据了大量的耕作时间,使得抓拍图像中监测范围位于农田有效区域的不足 40%。这些无效图像若不及时清除,不仅会占用网络传输通道和服务器存储空间,还将与有效图像一同作为秸秆覆盖率识别的样本基础和秸秆还田补贴的重要依据,大大降低对地块秸秆覆盖率计算的准确性。数据清洗[2]是降低数据冗余[3]和提高数据质量的关键步骤,也是减轻监测设备缓存压力和网络传输拥堵的有效环节。因此,准确高效的数据清洗方法[4]对推进智能化监测系统水平具有重要作用。

目前,时间序列清洗的研究一般采用单一条件时间维度的相关性来清洗一条时间序列,对多维时间序列的清洗研究较少。在单维度时间序列清洗方面,ZHANG 等[5]在基于平滑算法的外生变量的自回归模型基础上,提出了利用局部信息标注的迭代式修复数据清洗框架,修复检测到的异常值。TIAN 等[6]在基于统计的高斯混合模型基础上,提出了利用粒子群和最速下降法优化模型参数的时空概率模型,检测车辆燃油消耗量。YIN 等[7]基于速度约束的检测方法,利用加权平均值的方差约束检测速度时间序列中的异常值。SONG 等[8]提出了基于数据采集间隔的时间约束模型,解决数据传输过程中时间戳错乱问题。王咏梅等[9]针对数据仓库中模糊重复记录数据,提出基于 N-gram 的改进算法利用一个固定尺寸的优先队列窗口进行聚类,同时引入转换关闭准则检测相似重复记录。针对多维时间序列的预测方法[10],学者们发现使用机器学习方法能有效利用多维度信息[11-23]。上述研究对单一维度关系型数据的清洗方法较为系统,只有部分可以直接作用于时间序列数据,但图像抓拍系统中无效数据占比重较大,单一维度的时间序列无法实现较高拟合精度,而现有的多维时间序列方法没有很好利用序列间的物理特质,例如车速与时间约束了面积数据,电压会影响电流等。为此,本文对免耕播种机作业工况参数和抓拍图像数据进行研究,提出一种基于多条件时间序列的监测数据清洗方法,通过预测工况参数状态判断抓拍图像的清洗状态,从而用于数据后续的预测与决策。

1 数据与方法

1.1 研究区域与数据来源

本文选取位于松辽平原腹地的吉林省大安市作为研究区域,该地区以平原为主,地势平坦开阔,耕地总面积 135 400 hm^2,是世界著名的黄金玉米带。2020 年,大安市作为全省保护性耕作整体推进县,安装并投入使用免耕播种机及远程监测设备 880 余台,完成作业面积 72 000 hm^2,2020 年吉林省大安市免耕播种图层如图 1 所示。

图 1 2020 年吉林省大安市保护性耕作图层

Fig. 1 No-tillage layer in Da'an City in 2020

CTOPAM-DA2020 是 2020 年 4 月 15 日—6 月 20 日大安市保护性耕作免耕播种机作业参数数据集,包含了 15 000 多条免耕播种机作业真实的工况参数信息和 50 000 多幅秸秆覆盖图像。这些工况参数信息通过不同类型传感器每间隔 5 s 采集得到,包括免耕播种机位置、车速、瞬时面积、播种量等,秸秆图像由车载摄像头每 3 min 定时抓拍获取,包含不同免耕播种机不同时间的农田场景。由于数据量较庞大,模型训练需要大量人工标注工作,因此本文随机选取 5 条包含 3 000 个连续采样点的工况参数数据和在该时段内抓拍的 265 幅图像作为模型驱动的基础,图 2 为前 500 个采样点的变化趋势。

1.2 基于 LSTM 的工况参数训练

免耕播种机监测数据是一组长时间周期变化时

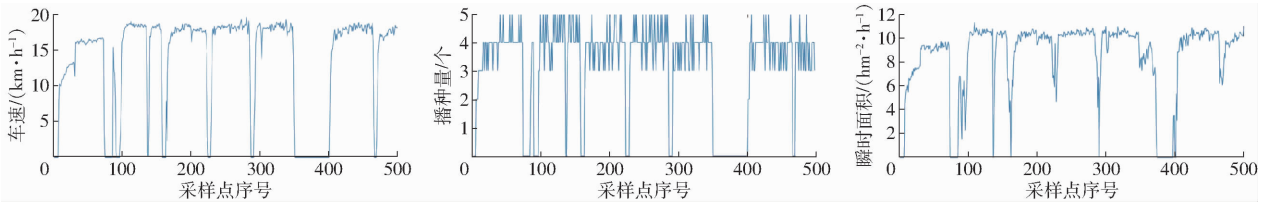

图 2　车速、播种量、瞬时面积时间序列(前 500 个点)

Fig. 2　Time series of vehicle speed, seeding rate and instantaneous area

间序列,当前的位置作业状态与历史工况信息相互关联。本文需要利用历史工况信息预测当前时刻的作业状态,因此需用到循环神经网络[24](Recurrent neural network,RNN)将信息向后传递。但随着信息过长时,历史信息的有效性越来越低,因此为避免长依赖问题长短时记忆[23](Long-short term memory,LSTM)的概念被提出,这是一种基于门控制的全互连循环神经网络,其网络结构如图 3 所示。由于当输入的工况时间序列相对较长时,LSTM 可通过每一个时间单元中的 4 个信息处理环节(丢弃、输入、存储、输出)来控制信息传递,从而建立较长时间的时序依赖关系,有效缓解 RNN 在长时序的梯度消失问题。因此,利用 LSTM 可以自动获取免耕播种机工况参数序列的时空特征,对工况状态进行准确预测。

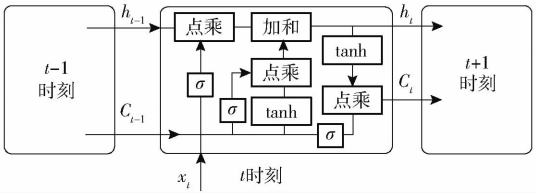

图 3　LSTM 网络结构

Fig. 3　Structure of LSTM network

以播种机工况参数中的车速时间序列为例,假设车速的时间序列样本为 $A=\{x_1,x_2,\cdots,x_n\}$,样本标签为 $Y=\{y_1,y_2,\cdots,y_m\}$,其中某一批次的子序列样本为 $X=\{x_1,x_2,\cdots,x_m\}$(m 远小于 n),则在 t 时刻的 LSTM 时间单元输入有:该单元的子序列输入信息 x_t,$t-1$ 时刻 LSTM 单元的输出信息 h_{t-1},$t-1$ 时刻 LSTM 单元的融合信息 C_{t-1};该单元的输出包括:t 时刻 LSTM 单元的输出信息 h_t,t 时刻 LSTM 单元的融合信息 C_t。

t 时刻的车速样本 x_t 进入网络后,其传递过程分别需经过 t 时刻丢弃部分信息 f_t、$t-1$ 时刻和 t 时刻融合信息 C_t、以及 t 时刻输出信息 h_t。其中 t 时刻的融合信息由输入信息 i_t 决定,而融合后的输出信息由 o_t 决定,因此其车速样本 x_t 前向推导公式为

$$f_t=\sigma(w_{fh}h_{t-1}+w_{fx}x_t+b_f) \tag{1}$$

$$i_t=\sigma(w_{ih}h_{t-1}+w_{ix}x_t+b_i) \tag{2}$$

$$C_t=f_tC_{t-1}+i_t\tanh(w_{ch}h_{t-1}+w_{cx}x_t+b_c) \tag{3}$$

$$o_t=\sigma(w_{oh}h_{t-1}+w_{ox}x_t+b_o) \tag{4}$$

$$h_t=o_t\tanh(C_t) \tag{5}$$

式中　w——权重,其中 $w=0$ 代表完全舍弃,$w=1$ 代表全部保留

b——空间映射中的偏移常量(Bias)

σ——sigmoid 激活函数

tanh——激活函数

由前向推导过程可以看出,t 时刻输出的车速信息包含了前一时刻车速信息以及当前时刻的新输入,而且模型对信息的选择具有随机性。

权重 w 的更新使用后向传播算法,由于输入车速样本 X 具有时序特性,则该方法演变为随时间的反向传播算法。t 时刻的 2 个残差分别为 δ_{ct} 和 δ_{ht},J_t 是以均方误差定义的损失函数,$\hat{y}_t$ 和 y_t 分别为 t 时刻免耕播种机的状态估计值和状态真值,则残差 δ_{ht} 和 δ_{ct} 的计算公式为

$$\delta_{ht}=\frac{\partial J_t}{\partial h_t}=\boldsymbol{W}^{\mathrm{T}}(\hat{y}_t-y_t)+\delta_{h,t+1}\frac{\partial h_{t+1}}{\partial h_t} \tag{6}$$

$$\delta_{ct}=\frac{\partial J_t}{\partial C_t}=\delta_{c,t+1}f_{t+1}+\delta_{ht}o_t(1-\tanh^2(C_t)) \tag{7}$$

式中　$\boldsymbol{W}$——权重 w 组成的权重矩阵

由式(6)和(7)的推导结果可以看出,由于特征信息 h_t 是由 x_t、C_t、o_t 共同控制的,在迭代过程中 δ_{ht} 的 $\frac{\partial h_{t+1}}{\partial h_t}$ 项不会总小于 1,而 δ_{ct} 是由后一刻的误差 $\delta_{c,t+1}$ 和本层的误差 δ_{ht} 组成,所以该结构不会发生梯度消失的问题。因此,将 LSTM 的时间单元在时间维度上展开,输入车速样本 $X=\{x_1,x_2,\cdots,x_m\}$ 的每一个时刻对应一个时间单元,即可以组成 m 层的 LSTM 特征提取模型,随着子序列步长 m 的变化,网络的层数可以无限深,经过激活函数 Softmax 参数优化后,在最后一层得到特征输出序列 $H=\{h_1,h_2,\cdots,h_m\}$。

2　多条件时间序列的免耕播种机监测数据清洗

保护性耕作监测设备在作业过程中每 3 min 就会传回 1 幅田间监测图像,每 5 s 采集一次工况参数,免耕播种机在田间往返作业时,工况参数幅

值会随时间呈周期性变化,而在无效区间内的数据又可分为边界和静止 2 种状态。其中,处于边界状态时车速和瞬时面积会在 20 ~ 30 s(即 4 ~ 6 个采集点)的范围内突然降低至零或缓慢增至正常车速,而静止状态时工况参数较长时间保持不变。车辆调头会占用大量时间,使得图像的抓拍位置通常位于地块两端、田间或地头静止。这类无效数据往往实时生成并实时传输,无法预先通过农田边界进行判定,若不及时清除会占用大量内存,并影响后续地块秸秆覆盖质量评价的准确性,因此本文通过某时刻免耕播种机工况参数的状态预测来间接实现图像清洗。

2.1 基于多条件时间序列的工况状态预测

由于农田长度限制和免耕播种机作业特性,免耕播种机在作业过程中保持较高的车速,行驶到农田两端会减速调头,并在农田间作往复运动,因此工况参数幅值会随时间呈周期性变化,如车速、播种量和瞬时面积。保护性耕作的有效监测范围为距离两端地头 10 m 的区间内,单一工况条件无法作为有效工况状态的判断依据,因此,搭建一个基于多条件时间序列的免耕播种机监测数据清洗模型,其网络结构如图 4 所示。

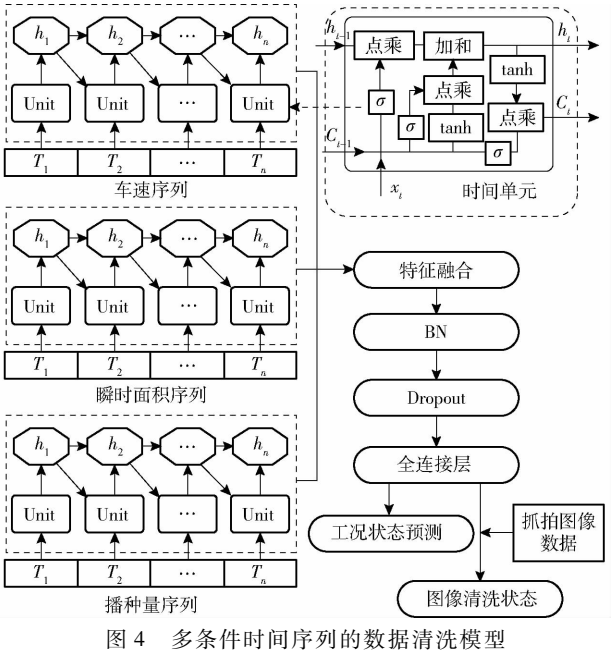

图 4 多条件时间序列的数据清洗模型

Fig. 4 Data cleaning model of multi-condition time series

(1)多条件时空特征提取

该网络结构的特征提取部分主要由 3 个单独的 LSTM 网络组成,分别对车速、播种量和瞬时面积进行时间特征提取。车速、播种量和瞬时面积的时间序列作为单个 LSTM 的网络输入,又将分成多个子序列,划分的步长作为影响模型预测性能的关键参数,将在实验部分详细讨论。

(2)多特征通道融合

车速、播种量和瞬时面积分别通过 LSTM 的时空特征提取后,得到的最后一个时刻的特征输出可以作为整个工况参数时间序列信息的融合。由于车速、播种量和瞬时面积对工况状态的响应不同,在融合前将每个特征乘以一个权重,可以提高不同条件的区分度,分别反映 3 种条件对预测结果的重要性。之后对 3 类特征进行通道融合,因为加减运算或平均求和的特征融合方法,都会使信息丢失,因此此处并不适合特征叠加。

(3)批量标准化位置后移

多特征通道融合后使得特征间干扰降低,批量标准化(Batch normalization,BN)可以使得网络中每个批次的特征输入保持较为标准的正态分布,但等比例的变化和偏移后不会造成信息缺失。由于本实验用到播种机的工况参数均在较短周期内循环,使得步长设定不宜过大,在特征提取之前加入 BN层容易阻碍信息传播,因此 BN 更适合放在特征融合之后。

(4)抑制过拟合

针对特征融合后的单一模型过拟合问题,需要在 BN 之后加入 Dropout 层,使得整个训练过程随着批次和迭代次数的增加更加复杂多样。在单次迭代中,某些网络可能会出现过拟合现象,但最后的输出结果取平均后,可以在一定程度上减轻整体网络的过拟合问题。

(5)数据清洗

模型通过分类器进行工况状态预测结果优化模型参数,最后将图像抓拍的时刻最近的工况参数代入训练好的模型中,根据工况预测结果清除落在无效时间区间内抓拍的图像,完成数据清洗。

2.2 基于作业参数的有效图像概率计算方式

针对免耕播种机作业监测系统图像抓拍的主要缺陷及其原因,首先从统计学与概率论的角度分析,分别统计得出了 200 组免耕播种机在田间作业单方向行驶 280 ~ 320 m 所用的平均时间为 2.25 s,完成一个作业周期用时平均时间为 7.75 min,从而可以得知车辆调头需用时 100 s 左右。因此,给出抓拍图像落在有效区间内的理论概率不超过 57% 。

本文将农田靠近两端 10 m 设为无效区域,地块中间区域定为有效区域,有效区域示意图如图 5 所示,图中图像标识为抓拍的位置点,绿色和灰色点为工况采集点,通过查找计数的方式,统计有效区域内抓拍图像的实际概率为

$$P_{AP}=\frac{N_A}{N}\times 100\% \tag{8}$$

式中 N_A——有效区域内图像标识的个数

N——整个地块内图像标识的个数

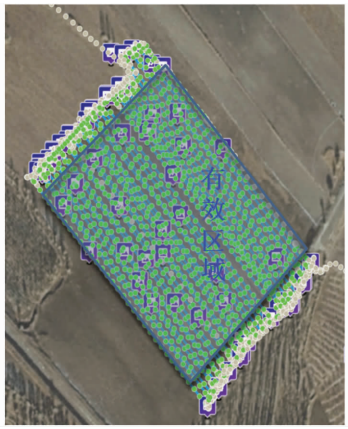

图 5　抓拍图像有效区域示意图

Fig. 5　Diagram for effective area of capture images

为了进一步探究在真实作业环境下免耕播种机抓拍图像的有效概率，并作为数据清洗结果的对比依据，研究了基于作业参数的抓拍有效图像概率计算方式。通过控制终端的数据库从中获取车辆当日作业的入场时间、出场时间、有效区间内的平均车速、农田面积以及垄长，则有效图像的概率可以通过有效时间和总时间的比值得到，即

$$E_P=\frac{1\,000M\left(1-\frac{10}{L}\right)^2}{(T_o-T_i)V_{ave}}\times 100\% \tag{9}$$

式中　M——地块作业总面积，hm^2

L——地块垄长，m

T_o——作业结束时刻

T_i——作业开始时刻

V_{ave}——平均车速，km/h

3　实验验证与分析

3.1　步长与分割步数选择

单个条件参数输入时间序列的步长与单个 LSTM 网络的层数一致，即 LSTM 网络每个时间单元对应每个时间子序列的输入值，是决定模型性能的关键因素。本文选取有代表性的 4 种分割方式((20,150)、(15,200)、(10,300)、(5,600)，其中前面的数字代表步长，后面的数字表示步数，迭代 40 次的准确率变化曲线如图 6 所示。

从图 6 中看出，将 3 000 维的时间序列分割成步长为 10，步数为 300 的子序列输入时训练的网络预测准确率最高，用时最少。当分割步长过长，即每个输入子序列过长，那么输入序列的时间信息丢失严重。当步长为 3 000 时，该网络等同于 BP 神经网络，无法提取时间特征。而当分割步长过短，即每个输入子序列过短，那么在时间信息传递过程中，前后时序相关性变弱，而且分割步数的增加，使得网络及

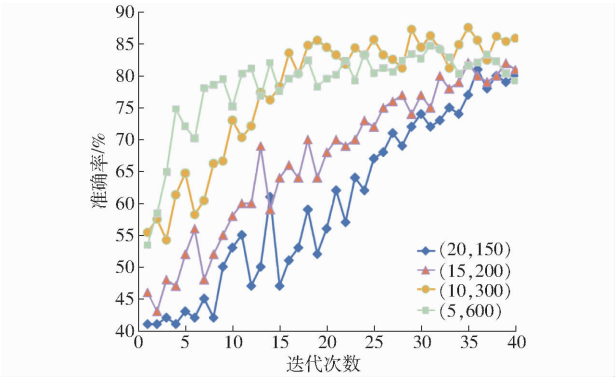

图 6　不同步长与步数的结果对比

Fig. 6　Results with different step-sizes and step-numbers

训练参数增加，收敛速度也会变慢。因此，本实验网络输入步长与分割步数取为(10,300)。

3.2　多条件特征融合对比

本文提出的模型最重要的改进点是利用多个 LSTM 分别对 3 种反映免耕播种机状态的工况参数进行时间特征提取，并且在通道融合时加入权重，使其在特征融合后能有所侧重，得到更能反映故障特性的融合层特征，从而提升网络故障诊断性能。多条件特征融合的状态预测模型训练过程准确率变化情况对比见图 7。由图 7 可知，多条件通道融合的 LSTM 的网络收敛更快，最终的预测精度也更高。

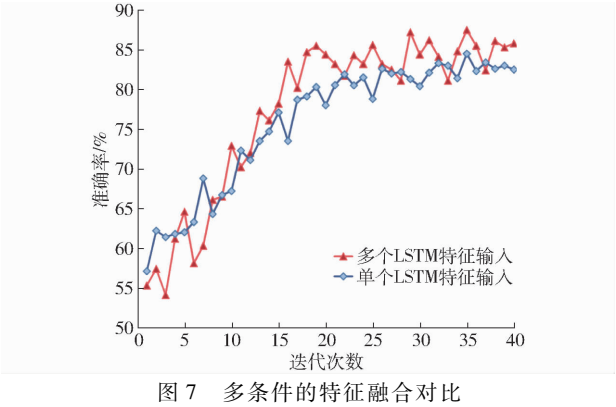

图 7　多条件的特征融合对比

Fig. 7　Comparison of multi-source input and only current input

3.3　不同模型测试结果

常用的状态预测网络框架有 BP 神经网络[26]、堆栈自动编码机[27]、循环神经网络等，在参数迭代 40 次的条件下实验结果对比如图 8 所示，分类器均使用 Softmax 函数。其中，BP 神经网络、堆栈自动编码机是基于条件参数序列的，其网络的神经元相对独立，没有提取空间和时间特征的能力，因此预测效果较差，无法挖掘出与时间关联性较强的不同工况状态特征。而基于 RNN 的算法由于在反向传播时，梯度更新项中激活函数导数的迭代难以避免地造成梯度弥散、信息丢失，所以在处理这些长序列时效果并不理想。本文提出的基于多条件时间序列预测模

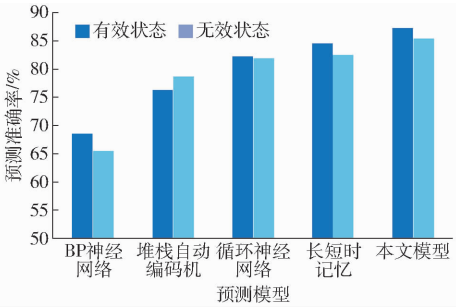

图 8 40 次迭代后不同预测模型结果对比

Fig. 8 Comparison of model results of different prediction networks after 40 iterations

型,不仅能提取输入序列时间特征,还能有效避免梯度消失问题,并且利用带权重层的融合层对 3 个 LSTM 提取的特征进行融合,有效地挖掘免耕播种机状态类型的特征,40 次迭代后对有效状态和无效状态的预测准确率都超过 85%。

3.4 基于作业参数的有效图像概率计算结果

将随机挑选的 5 个样本所在的当日作业数据代入式(8),得到概率计算结果如表 1 所示。从与真实值的对比结果可以看出,本文提出的有效概率均值为 33.15%,与实际概率均值 31.18% 差距较小,该方法的应用可省去大量人工计数的时间。从不同样本的计算概率可以看出,地块数量对有效图像概率有较大影响,地块数量越多,单位时间内作业的有效面积越少,抓拍图像位于无效区域的概率越大,使得落在有效区间内的平均概率远小于理论值(57%)。

另外,由于田间作业的地块数量和作业时长受驾驶员的主观意识和工作习惯的影响较大。当作业面积较小但地块较多时,农机手在田间转场和停止的时间明显增多,加大了抓拍图像落在无效区域的概率,并影响有效图像概率计算误差。因此较大的作业面积和较少的地块数量对有概率的计算精度影响较小。

表 1 有效图像概率计算结果

Tab. 1 Calculation results of effective picture probability

作业参数	样本 1	样本 2	样本 3	样本 4	样本 5
作业面积/hm^2	12.6	8.8	6.4	6.2	8.8
垄长/m	320	270	300	200	260
地块数量	2	2	3	2	1
车速/($km \cdot h^{-1}$)	15.3	14.7	15.1	14.3	14.9
作业时长/min	216	168	152	123	140
有效图像概率/%	35.7	33.1	26.1	31.7	39.2
实际概率/%	34.7	32.1	22.0	29.3	41.3

3.5 抓拍图像数据清洗实验结果

利用之前训练好的模型,将样本中与抓拍图像的时间最近的工况状态预测结果作为数据清洗的判断依据,得到的实验结果如表 2 所示。其中,清洗准确率为清洗数据中判断正确的样本占比,冗余率为抓拍样本中无效数据的占比,该部分应在采集时清洗。

表 2 抓拍图像清洗实验结果

Tab. 2 Results of data cleaning

清洗结果	样本 1	样本 2	样本 3	样本 4	样本 5
抓拍总数	72	56	50	41	46
有效数	24	17	9	12	20
清洗数	44	36	36	26	25
误判/漏判数量	2/4	2/3	3/5	3/3	2/1
准确率/%	95.5	94.4	91.7	88.5	92.0
冗余率/%	61.1	64.3	72.0	63.4	54.3

由表 1 和表 2 可以得出,表 2 中抓拍图像清洗平均准确率为 92.4%,5 类样本的误判或漏判结果较为均匀,数据清洗后约有 63% 的冗余数据被剔除,数据的存储和传输压力大幅降低。因此,利用本文的研究方法,以免耕播种机工况状态作为抓拍图像清洗依据是有效的。另外,结合表 1 的样本信息可以看出,地块数量较多和作业时长较短的样本对清洗准确率的影响较大,地块数量较多而面积较小的样本冗余率更高。

4 结论

(1)提出了基于多条件时间序列的免耕播种机监测数据清洗方法,该方法融合了多条件的 LSTM 特征提取和特征通道融合结构,通过预测免耕播种机工况状态,间接实现了抓拍图像清洗。同时,基于免耕播种机作业参数提出了一种有效图像概率的计算方式,该结果可代替人工计数的方式,作为抓拍图像清洗的参考数据。

(2)实验结果表明,对于不同地块数量、垄长、作业面积和时间的样本,基于多条件时间序列免耕播种机监测数据清洗方法,抓拍图像清洗平均准确率为 92.4%,数据清洗后约有 63% 的冗余数据被剔除。

(3)在免耕播种监测条件下,基于多条件时间序列的数据清洗方法可以准确预测播种机的工况状态和抓拍图像清洗状态,整体方法优于传统深度模型和单条件时序模型。该方法为免耕播种监测数据清洗及秸秆覆盖率的定量计算提供了技术支撑。

参 考 文 献

[1] 薛钦峰. 耕作有“利器”丰产有底气——黑土地上兴起保护性耕作[EB/OL]. [2020-06-11],中华人民共和国中央人

民政府网. http://www.gov.cn/xinwen/2020-06/11/content_5518552.htm
[2] CHU X, ILYAS I F, KRISHNAN S, et al. Data cleaning: overview and emerging challenges[C]// International Conference on Management of Data. ACM, 2016.
[3] ELMAGARMID A K, IPEIROTIS P G, VERYKIOS V S. Duplicate record detection: a survey[J]. IEEE Transactions on Knowledge & Data Engineering, 2007, 19(1):1-16.
[4] WANG X, WANG C. Time series data cleaning: a survey[J]. IEEE Access, 2020(8):1866-1881.
[5] ZHANG A, SONG S, WANG J, et al. Time series data cleaning: from anomaly detection to anomaly repairing[J]. Proceedings of the Vldb Endowment, 2017, 10(10):1046-1057.
[6] TIAN D, ZHU Y, DUAN X, et al. An effective fuel-level data cleaning and repairing method for vehicle monitor platform[J]. IEEE Transactions on Industrial Informatics, 2018, 15(1):410-422.
[7] YIN W, YUE T, WANG H, et al. Time series cleaning under variance constraints[C]// International Conference on Database System for Advanced Applications. Springer. Cham, 2018:108-113.
[8] SONG S,CAO Y,WANG J. Cleaning timestamps with temporal constraints[J]. PVLDB, 2016, 9(10):708-719.
[9] 王咏梅,陈家琪,嵇晓. 数据清洗中相似重复记录的识别[J]. 计算机科学与实践,2004,2(6):100-104.
WANG Yongmei, CHEN Jiaqi, JI Xiao. Detecting approximately duplicate database records in data cleaning[J]. Computer Science and Practice, 2004,2(6):100-104. (in Chinese)
[10] 胡珉,白雪,徐伟,等. 多维时间序列异常检测算法综述[J]. 计算机应用,2020,40(6):1553-1564.
HU Min, BAI Xue, XU Wei, et al. Review of anomaly detection algorithms for multidimensional time series[J]. Journal of Computer Applications, 2020,40(6):1553-1564. (in Chinese)
[11] MADAN R,SARATHIMANGIPUDI P. Predicting computer network traffic:a time series forecasting approach using DWT, ARIMA and RNN[C]// Proceedings of the IEEE 2018 Eleventh International Conference on Contemporary Computing, 2018:1-5.
[12] HADI A F, PURNOMO K. The ensemble of arima and gstar models in forecasting rainfall using Kalman filter[J]. International Journal of Scientific & Technology Research, 2020, 9(3):3640-3642.
[13] SAHED A. Using a GMDH-type neural network and ARIMA model to forecasting GDP in algeria during the period of 1990—2019[J]. American Journal of Economics and Business Management, 2020, 3(4):37-47.
[14] 吴静,李振波,朱玲,等.融合 ARIMA 模型和 GAWNN 的溶解氧含量预测方法[J].农业机械学报,2017,48(增刊):205-210.
WU Jing, LI Zhenbo, ZHU Ling, et al. Hybrid model of ARIMA model and GAWNN for dissolved oxygen content prediction[J]. Transactions of the Chinese Society for Agricultural Machinery, 2017, 48(Supp.):205-210. (in Chinese)
[15] WENG C Y. DWT-based reversible information hiding scheme using prediction-error-expansion in multimedia images[J]. Peer-to-Peer Networking and Applications, 2020,13:514-523.
[16] SHIVAM B, CHANDRASEKHAR E, PRIYANKA P, et al. A comparative study of wavelet-based ANN and classical techniques for geophysical time-series forecasting[J]. Computers & Geosciences, 2020, 138:104461.
[17] 陈英义,方晓敏,梅思远,等.基于 WT-CNN-LSTM 的溶解氧含量预测模型[J].农业机械学报,2020,51(10):284-291.
CHEN Yingyi, FANG Xiaomin, MEI Siyuan, et al. Dissolved oxygen prediction model based on WT-CNN-LSTM[J]. Transactions of the Chinese Society for Agricultural Machinery, 2020, 51(10):284-291. (in Chinese)
[18] JITENDRA KUMAR A, RIMSHA GOOMER B, ASHUTOSH KUMAR SINGH A. Long short term memory recurrent neural network (LSTM-RNN) based workload forecasting model for cloud datacenters[J]. Procedia Computer Science, 2018, 125:676-682.
[19] 温昌凯,谢斌,李若晨,等. 拖拉机作业载荷数据平台设计与旋耕作业质量预测[J]. 农业机械学报,2020,51(8):372-381.
WEN Changkai, XIE Bin, LI Ruochen, et al. Construction of tractor working load data platform and prediction of rotary tillage quality[J]. Transactions of the Chinese Society for Agricultural Machinery, 2020, 51(8):372-381. (in Chinese)
[20] YANG J, TAN K, MAT S, et al. Building energy consumption raw data forecasting using data cleaning and deep recurrent neural networks[J]. Buildings, 2019, 9(9):204.
[21] LIU C H,YANG L,DENG H,et al. Prediction of ammonia concentration in pigsty based on ARIMA and BP neural network[J]. Environmental Science in China,2019,39(6):2320-2327.
[22] LI D, CHEN D, JIN B, et al. MAD-GAN: multivariate anomaly detection for time series data with generative adversarial networks[C]// International Conference on Artificial Neural Networks, 2019: 703-716.
[23] 丁小欧,于晟健,王沐贤,等. 基于相关性分析的工业时序数据异常检测[J]. 软件学报,2020,31(3):726-747.
DING Xiaoou, YU Shengjian, WANG Muxian, et al. Anomaly detection on industrial time series based on correlation analysis[J]. Journal of Software, 2020,31(3):726-747. (in Chinese)
[24] MIAO Y, GOWAYYED M, METZE F. EESEN: End-to-End speech recognition using deep RNN models and WFST-based decoding[C]// IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), 2015:167-174.
[25] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8):1735-1780.
[26] GEHRING J, MIAO Y, METZE F, et al. Extracting deep bottleneck features using stacked auto-encoders[C]// IEEE International Conference on Acoustics, Speech and Signal Procession, 2013:3377-3381.
[27] ASL F H, SAEB M R, JAFARI S H, et al. Looking back to interfacial tension prediction in the compatibilized polymer blends: discrepancies between theories and experiments[J]. Journal of Applied Polymer Science, 2018, 135(16):46144.