第 37 卷 第 5 期
2022 年 10 月

天津科技大学学报
Journal of Tianjin University of Science & Technology

Vol. 37 No. 5
Oct. 2022

DOI:10.13364/j.issn.1672-6510.20220025

基于 Apriori 算法和卷积神经网络的风电机组故障诊断模型

张李炜，李孝忠
（天津科技大学人工智能学院，天津 300457）

摘 要：随着风能设备的规模在各国不断增大，风电机组的运行与维护成为研究热点. 针对风电机组的故障诊断问题，本文提出了一种基于 Apriori 算法和卷积神经网络（convolutional neural networks，CNN）的故障诊断模型. 该模型将 k 均值聚类（k-means）与 Apriori 算法结合进行特征选取并进行验证，以降低对专家经验的依赖性；以卷积神经网络构建故障诊断模型. 以真实风电场 SCADA（supervisory control and data acquisition）数据进行实验，通过准确率、精准率等指标将本文模型与其他模型进行对比. 结果表明，与其他模型相比，本文模型的准确率更高，整体效果更好.

关键词：Apriori 算法；卷积神经网络；SCADA 数据；故障诊断
中图分类号：TP391 **文献标志码**：A **文章编号**：1672-6510（2022）05-0050-06

Fault Diagnosis Model of Wind Turbine Based on Apriori and Convolutional Neural Network

ZHANG Liwei，LI Xiaozhong
（College of Artificial Intelligence，Tianjin University of Science & Technology，Tianjin 300457，China）

Abstract：With the increasing scale of wind energy equipment in various countries，the operation and maintenance of wind turbine has become a research hotspot. In this article，we propose a fault diagnosis model based on Apriori algorithm and convolutional neural network（CNN）to address the problem of fault diagnosis of wind turbine. This model combines k-means and Apriori algorithm for feature selection and verification so as to reduce the dependence on expert experience；the fault diagnosis model is constructed by CNN. The experiment was conducted with supervisory control and data acquisition（SCADA）data of real wind farm，and our proposed model was compared with other models including accuracy，accuracy and other indicators. The results show that compared with other models，the accuracy of this model is higher and the overall effect is better.

Key words：Apriori algorithm；convolutional neural network；SCADA data；fault diagnosis

近些年，随着新能源的热度越来越高，世界各国都在积极发展绿色能源. 风能作为规模较为成熟、产业链较为完善的绿色能源，具有能源储量大、开发成本相对较低的优势，因此全球风电机组的装机容量逐年增加. 但是，由于工作环境恶劣、工况多变的原因，风电机组在运行中后期故障频发，维护成本也不断增加，所以做好风电机组的故障诊断工作可以有效降低运营和维护的成本，提高机组运行效率[1-2].

目前，风电机组故障诊断的方法主要为基于解析模型与基于数据驱动的故障诊断方法[3]. 基于解析模型的故障诊断方法是早期的研究方向，由于早期传感器还不够成熟，不能获取机组的准确状态，只能通过研究机组本身的物理模型来进行故障诊断. 随着大数据、人工智能等新兴信息技术的发展，大量学者开始将重心转移到基于数据驱动的故障诊断研究中. 金晓航等[4]通过使用稀疏自编码神经网络对 SCADA（supervisory control and data acquisition）的数据进行模型训练与故障预测，能够使风电机组的异常状态在

收稿日期：2022-02-16；**修回日期**：2022-04-29
作者简介：张李炜（1998—），男，江苏人，硕士研究生；**通信作者**：李孝忠，教授，lixz@tust.edu.cn

出现初期被发现；但是，判定数据出现异常的阈值是人为设定的，而且根据经验设定的阈值也不总是准确的. 任建亭等[5]提出一种基于深度变分自编码网络的风电齿轮箱故障预警方法，通过深度变分自编码网络与 SCADA 数据结合，在学习数据结构特征的同时挖掘数据的分布规则；但是，输入到网络模型中的 SCADA 数据特征参数完全依靠人的先验知识，具有一定的主观性. Reder 等[6]意识到对 SCADA 数据进行特征选取的重要性，提出一种基于数据驱动的关联天气条件与风电涡轮机故障的学习框架，通过 Apriori 算法挖掘和解释 SCADA 特征参数与故障之间的相关性，但是没有将特征选取结果结合故障诊断模型进行诊断. 靳志杰等[7]提出一种基于特征选取的 XGBoost 风电机组故障诊断方法，比传统机器学习算法准确率更高；但是，其特征选取算法时间效率低下，即使只获取 6 个特征参数的重要指数，也需消耗近 4 h.

本文提出一种基于 Apriori 算法和卷积神经网络（convolutional neural networks，CNN）的风电机组故障诊断模型. 在该模型中，使用 k 均值聚类（k-means）算法对 SCADA 数据进行聚类，将聚类结果输入 Apriori 算法的关联规则算法中进行提升度排序，然后进行特征选取并使用模型验证，最后用卷积神经网络模型进行故障诊断. 使用爱尔兰某风电场真实 SCADA 数据进行实验，实验表明本文模型诊断效果比其他两种神经网络模型更好.

1　基于 k-means 算法与 Apriori 算法的特征选取

1.1　数据预处理

SCADA 是安装在大型风电机组上的一套用于监测风电机组运行状态的传感器系统. SCADA 监测的零部件包括发电机、主轴、齿轮箱、变流器、控制柜等，具体采集的数据为 100 多个离散特征数据（如某子系统或者零部件的状态、动作等）或者连续特征数据（如机组发电功率、转速等）.

第一步工作是将 SCADA 原始数据进行预处理，主要是将一些不合理数据进行剔除，例如在传输过程中产生的无效数据、在维修或者保养机组过程中产生的异常数据等[8]. 第二步工作是将筛选后的 SCADA 数据结合机组状态的运行数据进行故障状态标记.

1.2　特征选取

SCADA 系统采集的数据十分丰富，在针对某一故障进行分析时，仅仅依靠专家经验或者简单的相关性分析可能会遗漏相关性较弱的数据，或者是将冗余数据当作相关性数据，这些都将降低模型的准确性[9]. 因此，本文使用 k-means 算法与 Apriori 算法相结合，评估故障的相关性特征数据.

k-means 算法是一种迭代求解的无监督学习算法[10-11]，目的是将某一故障所有数据的每一个特征进行聚类，然后定义其阈值，最后将特征数据全部替换成该特征数据所处的类别，以便为下一步 Apriori 算法的运用提供合适的输入数据.

Apriori 算法是在一个数据集中发现项集之间的有趣关联或者相互联系[12]. 其评价指标如下：

（1）支持度计数：指包含特定项集事务的个数. 在数学上，项集 X 的支持度计数可以表示为 $\sigma(X)=|\{t_i \mid X \subseteq t_i, t_i \in T\}|$.

（2）支持度：项集 X 与项集 Y 同时发生的概率称为 $X \to Y$ 关联规则的支持度，可表示为 $s(X \to Y)=\sigma(X \cup Y)/N$.

（3）置信度：如果项集 X 发生，项集 Y 也发生的概率称为 $X \to Y$ 关联规则的置信度，可表示为 $c(X \to Y)=\sigma(X \cup Y)/\sigma N$.

（4）提升度：项集 X 发生对项集 Y 发生的概率产生多少变化称为 $X \to Y$ 的提升度，可表示为 $l(X \to Y)=c(X \cup Y)/\sigma(Y)$ [13].

Apriori 算法是使用一种逐层搜索的迭代方法，其目标是找出支持度大于等于最小支持度并且置信度大于等于最小置信度的所有规则，其中最小支持度与最小置信度分别对应支持度和置信度的阈值. 整体算法大致分为两步：

（1）频繁项集的产生：寻找满足最小支持度的所有项集，把这些项集称作频繁项集.

（2）规则的产生：从上一步频繁项集中提取满足高置信度的规则，把这些规则称作强规则.

2　故障诊断模型

2.1　卷积神经网络

卷积神经网络主要由输入层、隐藏层、全连接层和输出层组成，其中隐藏层主要是由卷积层、激活层和池化层组成[14].

2.1.1　卷积层

卷积层由多个卷积核组成，每个卷积核对输入数据进行扫描，进而提取不同的特征，这步操作即为

卷积. 卷积运算公式为

$$y_j^l = \sum_{i \in M} x_i^{l-1} * w_{ij}^l \tag{1}$$

其中：y_j^l 表示第 l 层第 j 个输出特征图，M 表示所有输入特征图的集合，x_i^{l-1} 表示第 $l-1$ 层第 i 个输入特征图，*表示卷积操作，w_{ij}^l 表示卷积核.

2.1.2　激活层

激活层是将卷积层的输出结果进行非线性映射. 卷积完成后，通常会在后面加入偏置，再引入非线性激活函数，这里偏置可表示为 b，激活函数可表示为 $g()$，然后得到经过激活函数运算的结果为

$$x_j^l = g(y_j^{l-1} + b_j) \tag{2}$$

CNN 一般选取 ReLU 函数作为激活函数.

2.1.3　池化层

特征图经过激活函数后进入池化层，主要操作为池化，是一种降采样操作. 池化层的输出可表示为

$$x_j^l = p(x_j^{l-1}) \tag{3}$$

其中 $p()$ 表示池化函数. 常用的池化操作有最大值池化法和平均值池化法.

CNN 常被用于图像识别领域，由于图像是二维数据，所以使用二维卷积神经网络；而本文处理的数据是在时间上有关联的序列数据，为一维数据，所以本文使用一维卷积神经网络. 一维卷积神经网络与二维卷积神经网络类似，都具有 CNN 局部连接与权值共享的特性，局部连接可以减少模型所需的参数，权值共享则能很好地避免模型过拟合[15].

2.2　建模流程

故障诊断模型构建流程如图 1 所示.

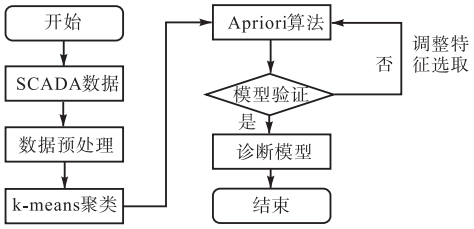

图 1　故障诊断模型构建流程

Fig. 1　Construction process of fault diagnosis model

本文基于 Apriori 算法和 CNN 实现风电机组的故障诊断，具体步骤如下：

(1) 数据预处理部分对采集的风电机组 SCADA 数据进行筛选、剔除，减少异常数据对后续实验的影响.

(2) 将预处理后的数据结合风电机组状态的运行数据进行状态标记.

(3) 选定某一故障，将该故障的所有数据按特征参数输入 k-means 算法进行聚类，然后将所有特征数据替换成聚类后的类别作为输出结果.

(4) 将步骤 (3) 的输出结果作为 Apriori 算法的输入，得到每个特征与该故障的相关度，然后结合诊断模型的准确性和消耗时间选择特征集.

(5) 按照特征集获取数据，再将数据划分训练样本与测试样本后输入诊断模型进行训练与测试.

3　实　验

3.1　数据集

实验所使用的数据为爱尔兰某风电场 2014 年 5 月 1 日至 2015 年 4 月 9 日的 SCADA 数据，其中包括 49 028 条 63 个特征的运行数据、27 398 条风能转换器 (WEC) 状态数据和 1 850 条远程终端单元 (RTU) 状态数据. WEC 对应与机组本身直接相关的状态数据，RTU 对应机组连接到电网的功率控制数据. 该数据集覆盖了风电机组的正常与异常运行状态. 实验分别对交流器馈电 (故障 A) 与发电机励磁 (故障 B) 两种故障进行诊断.

3.2　特征选取

使用 k-means 算法对两种故障的 27 个初步选取的特征进行聚类，初始的聚类中心从 3 个增加到 10 个，然后使用轮廓系数评判聚类效果. 将 k-means 的输出结果输入 Apriori 算法：首先进行独热编码；然后寻找频繁项集，输出最小支持度大于等于 0.5、最大长度为 2 的频繁项集；最后寻找强规则，输出最小置信度大于等于 0.7 的规则. 提升度反映特征变量与发生故障之间的相关性，故障 A 和故障 B 的提升度见表 1 和表 2.

表 1　故障 A 的提升度

Tab. 1　Lift of fault A

序号	特征	提升度
1	1 号转子温度	1.269 231
2	1 号定子温度	1.269 231
3	2 号定子温度	1.269 231
4	变桨柜叶片 A 温度	1.260 935
5	轮毂温度	1.260 935
6	塔筒温度	1.259 315
7	机舱温度	1.205 769
8	1 号机舱环境温度	1.198 280
9	2 号机舱环境温度	1.198 280
10	变压器温度	1.195 652
11	变桨柜叶片 B 温度	1.194 017
12	变流器柜风扇温度	1.181 361
13	2 号转子温度	1.063 601

续表

序号	特征	提升度
14	控制柜温度	1.057 692
15	桨距角	1.048 495
16	平均风速	1.037 221
17	前轴承温度	1.010 068
18	变桨柜叶片 C 温度	1.010 068
19	环境温度	1.004 808
20	整流柜温度	0.996 105
21	有功功率	0.995 750
22	后轴承温度	0.969 231

表 2　故障 B 的提升度

Tab. 2　Lift of fault B

序号	特征	提升度
1	控制柜温度	1.130 952
2	机舱机柜温度	1.130 952
3	变桨柜叶片 A 温度	1.130 952
4	变桨柜叶片 B 温度	1.130 952
5	变桨柜叶片 C 温度	1.130 952
6	整流柜温度	1.130 952
7	1 号转子温度	1.130 952
8	塔筒温度	1.130 952
9	变压器温度	1.130 952
10	桨距角	1.130 952
11	1 号定子温度	1.119 293
12	2 号定子温度	1.119 293
13	平均风速	1.078 104
14	后轴承温度	0.976 732
15	2 号转子温度	0.929 681
16	有功功率	0.923 611

由表 1 和表 2 可知：有 22 个特征与故障 A 之间是强规则关系，但只有 19 个特征的提升度大于 1，所以故障 A 的特征从这 19 个特征中选取；同样，有 16 个特征与故障 B 之间是强规则关系，但只有 13 个特征的提升度大于 1，所以故障 B 的特征从这 13 个特征中选取.

3.3　模型验证

使用 CNN 与长短期记忆（long short-term memory，LSTM）网络对上述所选的特征进行验证. 使用准确率、精准率、召回率和 F_1 值作为特征选取的评价指标，结果如图 2—图 5 所示.

由图 2 和图 3 可知：在 CNN 与 LSTM 两个模型中加入了第 20—第 22 个特征后，故障 A 在 LSTM 模型中的准确率、精准率和召回率出现了小幅度降低，而在 CNN 模型中的则基本保持不变. 这表明第 20—第 22 个特征与故障 A 相关度不高. 在 CNN 与 LSTM 两个模型中，各指标均在第 16 个特征之前呈增长趋势，在第 16 个特征之后趋于平缓，说明第 17—第 19 个特征带给模型的作用微乎其微，与其提升度相对应，第 17—第 19 个特征的提升度只比 1 高 0.01 左右，所以故障 A 选择第 1—第 16 个特征作为故障诊断模型的输入参数.

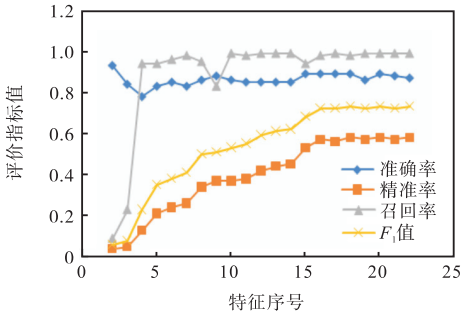

图 2　故障 A-CNN 的评价指标

Fig. 2　Evaluation indicator of fault A-CNN

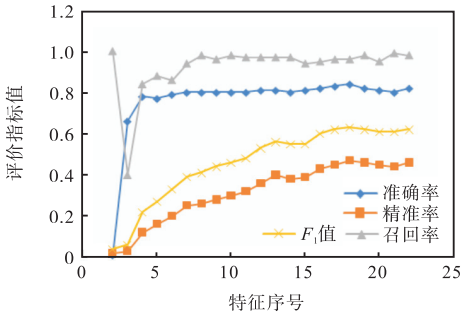

图 3　故障 A-LSTM 的评价指标

Fig. 3　Evaluation indicator of fault A-LSTM

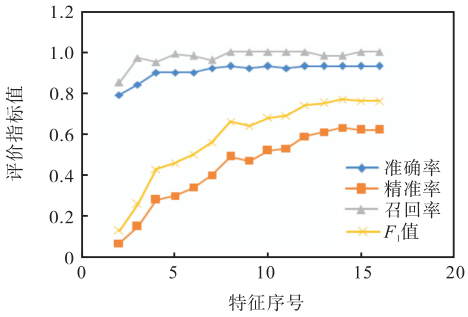

图 4　故障 B-CNN 的评价指标

Fig. 4　Evaluation indicator of fault B-CNN

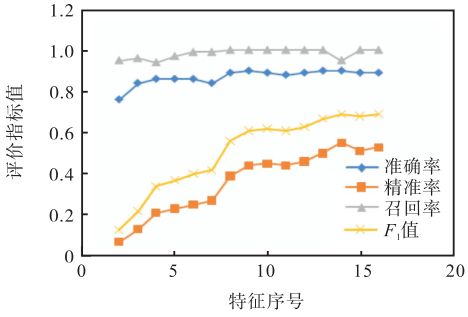

图 5　故障 B-LSTM 的评价指标

Fig. 5　Evaluation indicator of fault B-LSTM

由图 4 和图 5 可知：在 CNN 与 LSTM 两个模型中，故障 B 在加入了第 14—第 16 个特征之后，准确率与 F_1 值均趋于平缓，召回率与精准率出现小幅度变化，但总体趋于平缓，所以故障 B 还是按照提升度选取第 1—第 13 个特征作为模型参数.

3.4　模型诊断

根据上述验证结果，基于 Apriori 算法故障 A 选取第 1—第 16 个特征，故障 B 选取第 1—第 13 个特征，然后结合 CNN 模型进行故障诊断，并与 LSTM、多层感知器神经网络（multilayer perceptron，MLP）和支持向量机（support vector machine，SVM）模型的结果进行对比，结果见表 3.

在故障 A 中，CNN 模型的准确率、精准率与 F_1 值均为最高，仅召回率稍低于 SVM 模型. MLP 模型为单一的神经网络模型，利用线性层对时序数据进行信息提取较为困难，所以效果是 4 个模型中最差的. SVM 模型是 4 个模型中唯一的机器学习算法，综合前 4 个指标来看，其效果仅次于本文基于 Apriori 算法的 CNN 模型，但 SVM 模型的优势就是不用训练神经网络，所以耗时最短. LSTM 作为循环神经网络的代表，在准确率上差强人意但耗时最长.

表 3　故障诊断结果

Tab. 3　Results of fault diagnosis

故障类型	模型算法	准确率	精准率	召回率	F_1 值	耗时/s
故障 A	MLP	0.46	0.13	0.52	0.21	82.6
	SVM	0.85	0.48	0.99	0.64	42.8
	LSTM	0.82	0.43	0.95	0.60	767.1
	CNN	0.89	0.57	0.98	0.72	80.8
故障 B	MLP	0.46	0.1	0.54	0.17	52.6
	SVM	0.92	0.56	1.00	0.72	8.0
	LSTM	0.90	0.50	1.00	0.67	114.6
	CNN	0.93	0.61	0.98	0.75	49.9

在故障 B 中，CNN、LSTM、SVM 模型的准确率均不低于 0.9，其中 CNN 模型最高，同时精准率与 F_1 值同样最高. 与其他两个神经网络模型相比，CNN 模型在耗时上有明显的优势.

4　结　语

本文提出了一种基于 Apriori 算法和卷积神经网络的风电机组故障诊断模型，将 k-means 聚类算法与 Apriori 算法的关联规则算法结合起来进行特征选取，并使用两种神经网络模型进行验证，然后将选取后的特征运用到卷积神经网络模型中进行故障诊断. 最后采取真实风电场的 SCADA 数据进行实验，并使用另外两种神经网络和一种机器学习算法进行对比实验，实验结果表明本文方法效果最好，准确率最高. 但是，本文方法的精准率不高，说明该模型容易将正常运行的状态判定为故障状态，后续还需进一步提高模型的精准率.

参考文献：

[1] 郭莹莹，张磊，肖成，等. 基于改进深度森林算法的风电机组故障诊断技术研究[J]. 可再生能源，2019，37(11)：1720–1725.

[2] 孟宪梁，梁伟，杨志，等. 基于机器学习算法与 SCADA 系统的风电机组变桨系统变频器的故障预警方法研究[J]. 太阳能，2021(2)：78–84.

[3] 龙霞飞，杨苹，郭红霞，等. 大型风力发电机组故障诊断方法综述[J]. 电网技术，2017，41(11)：3480–3491.

[4] 金晓航，王宇，ZHANG B. 工业大数据驱动的故障预测与健康管理[J/OL]. 计算机集成制造系统：1–27[2022–01–31]. http://kns.cnki.net/kcms/detail/11.5946.TP.20200814.1703.006.html.

[5] 任建亭，汤宝平，雍彬，等. 基于深度变分自编码网络融合 SCADA 数据的风电齿轮箱故障预警[J]. 太阳能学报，2021，42(4)：403–408.

[6] REDER M，YÜRÜŞEN N Y，MELERO J J. Data-driven learning framework for associating weather conditions and wind turbine failures[J]. Reliability engineering & system safety，2018，169：554–569.

[7] 靳志杰，霍志红，许昌，等. 基于特征选择和 XGBoost 的风电机组故障诊断[J]. 可再生能源，2021，39(3)：353–358.

[8] 刘宁. 风电场运行数据的关联分析研究[D]. 天津：河北工业大学，2015.

[9] 邓子豪，李录平，刘瑞，等. 基于 SCADA 数据特征提取的风电机组偏航齿轮箱故障诊断方法研究[J]. 动力工程学报，2021，41(1)：43–50.

[10] 吴婷婷，李孝忠，刘徐洲. 基于 k-means 的改进协同过滤算法[J]. 天津科技大学学报，2021，36(6)：44–48.
[11] FAHIM A. K and starting means for k-means algorithm[J]. Journal of computational science，2021，55：101445.
[12] 黄小红，陈丽华，王倩. 基于改进 Apriori 算法的 SCADA 系统事故后数据分析[J]. 华北电力大学学报（自然科学版），2008(4)：27–32.
[13] TONG C，GUO P. Data mining with improved Apriori algorithm on wind generator alarm data[C]//IEEE. 2013 25th Chinese Control and Decision Conference(CCDC). New York：IEEE，2013：1936–1941.
[14] 黎阳羊，胡金磊，赖俊驹，等. 基于 1D-CNN-LSTM 混合神经网络模型的风电机组行星齿轮箱故障诊断[J]. 电气自动化，2021，43(5)：20–22.
[15] 李东东，王浩，杨帆，等. 基于一维卷积神经网络和 Soft-Max 分类器的风电机组行星齿轮箱故障检测[J]. 电机与控制应用，2018，45(6)：80–87.

责任编辑：郎婧

（上接第 43 页）

[3] SOLTANI A R，TAWFIK H，GOULERMAS J Y，et al. Path planning in construction sites：performance evaluation of the Dijkstra，A*，and GA search algorithms[J]. Advanced engineering informatics，2002，16(4)：291–303.
[4] MAC T T，COPOT C，TRAN D T，et al. Heuristic approaches in robot path planning：a survey[J]. Robotics & autonomous systems，2016，86：13–28.
[5] 彭澎. 基于 A*算法的路径规划算法研究[D]. 合肥：安徽工业大学，2018.
[6] 张儒珂. 基于改进人工势场法的自动超车控制方法研究[D]. 大连：大连理工大学，2021.
[7] 马子涵. 四足机器人机体姿态及运动控制算法仿真研究[D]. 西安：中国科学院大学（中国科学院西安光学精密机械研究所），2020.
[8] SINGH Y，SHARMA S，SUTTON R，et al. A constrained A* approach towards optimal path planning for an unmanned surface vehicle in a maritime environment containing dynamic obstacles and ocean currents[J]. Ocean engineering，2018，168(1)：187–201.
[9] VISWANATHAN P，ZAMBALDE E P，FOLEY G，et al. Intelligent wheelchair control strategies for older adults with cognitive impairment：user attitudes，needs，and preferences[J]. Autonomous robots，2017，41(3)：539–554.
[10] 杨冰. 一种适用于复杂地形的智能轮椅的设计[D]. 太原：中北大学，2021.
[11] TSENG P H，RAJANGAM S，LEHEW G，et al. Interbrain cortical synchronization encodes multiple aspects of social interactions in monkey pairs[J]. Scientific reports，2018，8(2)：177–192.
[12] KUCUKYILDIZ G，OCAK H，KARAKAYA S，et al. Design and implementation of a multi sensor based brain computer interface for a robotic wheelchair[J]. Journal of intelligent & robotic systems，2017，87(2)：247–263.
[13] SAIKIA A，HAZARIKA S M. cBDI：towards an architecture for human-machine collaboration[J]. International journal of social robotics，2017，9(2)：211–230.
[14] 张焕林. 自动驾驶轮椅沿路沿行驶的方法研究与实现[D]. 西安：西安电子科技大学，2020.
[15] 高晓阳. 基于改进人工势场法的自主机器人动态避障研究[D]. 郑州：郑州大学，2020.

责任编辑：郎婧