基于数据挖掘和人工神经网络的厌氧产气模型构建

温沁雪[1]， 李奕芯[2]， 杨 硕[1]， 党 宁[3]， 甘硕儒[1]， 李慧莉[3]， 陈志强[1,3]
（1. 哈尔滨工业大学 城市水资源与水环境国家重点实验室，黑龙江 哈尔滨 150090；2. 中国建筑西南设计研究院有限公司，四川 成都 610041；3. 兰州理工大学 土木工程学院，甘肃 兰州 730050）

摘 要： 为模拟厌氧产甲烷规律，以稳定运行的猪粪秸秆厌氧共消化反应器为基础，应用数据挖掘技术对反应器的运行数据进行预处理，并应用人工神经网络建立猪粪秸秆厌氧共消化产气预测模型。利用 SPSS 中的聚类分析和相关性分析，确定模型的输入、输出个数分别为 5 和 1，利用试凑法确定模型隐藏层个数为 9。确定模型拓扑结构后，应用标准 BP 算法和动量－自适应学习速率算法训练模型，结果表明动量－自适应学习速率算法有更好的训练效果。同时，对模型的性能进行验证发现，该模型的性能较好，说明厌氧消化产气预测模型具有一定的适用性。

关键词： 厌氧共消化； 人工神经网络； 数据挖掘； 猪粪； 秸秆； 产气预测模型
中图分类号： TU993.3 **文献标识码：** A **文章编号：** 1000－4602(2019)01－0077－05

Anaerobic Co-digestion Biogas Production Model Based on Data Mining and Artificial Neural Network

WEN Qin-xue[1]， LI Yi-xin[2]， YANG Shuo[1]， DANG Ning[3]， GAN Shuo-ru[1]，
LI Hui-li[3]， CHEN Zhi-qiang[1,3]
(1. *State Key Laboratory of Urban Water Resource and Environment*, *Harbin Institute of Technology*, *Harbin* 150090, *China*; 2. *China Southwest Architectural Design and Research Institute Co. Ltd.*, *Chengdu* 610041, *China*; 3. *School of Civil Engineering*, *Lanzhou University of Technology*, *Lanzhou* 730050, *China*)

Abstract: Stable operation data from anaerobic co-digestion reactor of pig manure and straw was used as raw data to simulate the anaerobic methane production process. Data mining was used for the pretreatment of data and the artificial neural network was used for the development of the biogas prediction model. Five inputs and one output of the model were determined using the cluster analysis and correlation analysis in SPSS. The number of neurons in the model was nine, selected using the method of cut-and-try. Two commonly used algorithms, standard BP algorithm and momentum-learning rate adaptive algorithm, were used to train the model after ensuring model topology. Results showed better performance in the momentum-learning rate adaptive algorithm. The model was tested and showed good performance, which indicated applicability of the methane production prediction model.

Key words: anaerobic co-digestion; artificial neural network; data mining; pig manure; straw; biogas production prediction model

基金项目：国家重点研发计划项目(2016YFC0401102－3)； 国家自然科学基金资助项目(51578182)
通信作者：陈志强 E－mail:czqhit@163.com

厌氧消化是常用的有机废物实现资源化的方法,但是传统的厌氧消化技术在处理单一基质,比如污泥、粪便时,存在产气量低的问题,而选择厌氧共消化处理污泥和粪便则可有效解决易发生氨抑制及甲烷产率低的问题[1]。近年来,利用数据挖掘进行建模被广泛应用于环境工程领域[2-4]。由于厌氧共消化具有复杂性和非线性等特点,因此对该过程进行建模是环境工程领域的研究难点。与厌氧消化相关的模型如 IPCC 模型、COD 估算模型等,具有参数较难获得、厌氧过程过于简化、误差较大等缺点[5-6],不能很好地描述全部厌氧消化过程和指导实际生产运行。人工神经网络通过学习逼近任意非线性映射,便于表征复杂成分及数量的变化,可以用于厌氧共消化的建模过程[7]。

笔者以稳定运行的猪粪秸秆厌氧共消化反应器的运行数据为依据,应用数据挖掘技术对其进行预处理,并应用人工神经网络建立猪粪秸秆厌氧共消化产气预测模型,旨在为模拟猪粪秸秆厌氧共消化过程提供依据,并指导规模化养殖过程中粪便和秸秆的处理问题。

1　材料与方法

1.1　试验材料

本试验所用的猪粪和秸秆均来自于哈尔滨市某养猪场。粪便去除杂质后密存,秸秆磨粉后备用(长约为 2 mm,宽约为 0.25 mm)。猪粪和秸秆的总固体含量分别为 23.5% 和 82.4%,挥发性固体含量分别为 17.4% 和 64.8%,碳含量分别为 46.26% 和 45.17%,氮含量分别为 4.25% 和 0.63%。污泥取自某玉米加工厂 EGSB 厌氧反应罐,随后以蔗糖和猪粪 - 秸秆混合物进料,逐步替换为猪粪 - 秸秆混合物。

试验过程中,控制反应器的温度为 35 ℃,反应体积为 5 L,SRT 为 20 d。控制进料中猪粪与秸秆的质量比为 2∶1,进料混合固体 TS 为 7%。

1.2　试验方法

检测指标包括 pH 值、ORP 值、挥发性脂肪酸(VFAs)、溶解性 COD(SCOD)、NH_4^+ - N、碱度、总产气量、甲烷产量、TS、VS 和 C/N 值。其中,pH 值与 ORP 值采用 pH 计和氧化还原电极测定,SCOD、氨氮、碱度和碳氮比分别采用消解 - 化学滴定法、分光光度法、电位滴定法和元素分析仪测定,VFAs 和 CH_4 采用气相色谱测定,TS 和 VS 采用重量法测定。猪粪秸秆厌氧反应器稳定运行了 90 d,排除错误数据后,共获得 83 组完整数据,将这些数据作为 BP 神经网络的原始数据。

2　结果与讨论

2.1　数据的相关性分析

本试验测定了多个常规指标,但这些指标与产气量不一定直接相关。因此,需要提前进行相关性分析,从而确定建模所需的指标。利用 SPSS 22.0 软件分析试验数据之间的相关性,并选用斯皮尔曼等级相关系数法分析 VFAs、碱度、氨氮、SCOD、前一天总产气量和 CH_4 含量与甲烷产量的相关性,结果见表 1。可以看出,氨氮与甲烷产量的相关性较为显著;VFAs、SCOD、CH_4 含量、前一天产气量与甲烷产量的相关性非常显著;碱度与甲烷产量的相关性不明显。这与赖夏颉的研究结果相似[7],说明本试验的结论具有一定的可信度。

表 1　厌氧发酵常规理化指标的相关系数

Tab. 1　Correlation coefficient of physicochemical indexes in anaerobic fermentation

项　目	甲烷产量	VFAs	SCOD	碱度	氨氮	前一天产气量	CH_4 含量
甲烷产量	1.00	-0.29**	-0.54**	-0.02	-0.24*	0.51**	0.67**
VFAs	-0.29**	1.00	0.30**	0.04	0.18	-0.25*	-0.13
SCOD	-0.54**	0.30**	1.00	0.06	0.41**	-0.42**	-0.50**
碱度	-0.02	0.04	0.06	1.00	0.10	-0.01	0.14
氨氮	-0.24*	0.18	0.41**	0.10	1.00	-0.06	-0.12
前一天产气量	0.51**	-0.25*	-0.42**	-0.01	-0.06	1.00	0.48**
CH_4 含量	0.67**	-0.13	-0.50**	0.14	-0.12	0.48**	1.00

注:　* 表示相关性显著,而 * * 表示相关性非常显著。

2.2　错误数据的剔除

测量出的试验数据由于多种原因,不可避免地会出现一些误差。为了提高模型的精度,需要去掉其中的错误数据。本试验拟采用 K 平均聚类算法筛选原始数据,设置初始聚类中心为 6,聚类结果如表 2 所示。1、2、3、4、5、6 聚类中包括的数据总数分

别为 22、6、26、1、20、8。可以看出,聚类中心 4 的数据仅有 1 个,小于总数据量的 1.5%,因此舍弃第 4 个聚类中心。之后利用欧几里得度量计算聚类中心与数据点之间的距离,经过计算之后,在收集到的数据中最终得到 81 组有效数据,采用该 81 组数据建立模型。

表 2　聚类结果

Tab. 2　Clustering results

项　目	聚类中心					
	1	2	3	4	5	6
甲烷产量	194.01	176.77	181.96	133.53	216.98	225.59
VFAs	19.26	30.18	21.47	1 211.69	17.99	13.09
SCOD	716.35	1 414.99	1 031.37	699.99	290.49	238.81
氨氮	646.72	784.18	841.69	571.98	556.59	890.87
前一天产气量	191.12	182.36	188.09	122.00	211.59	225.58
甲烷含量	54.22	53.33	53.01	50.02	57.02	57.49

2.3　模型的建立

为了均等对待每个指标,消除因单位和数值量级差距较大带来的影响,在建模前对不同指标进行归一化预处理。采用 mapminmax 函数处理数据,该函数是一种常见的归一化处理方法,可以把样本数据利用相应的算法归一化到最优的数据范围内。

2.3.1　隐含层层数与输入、输出层个数的确定

相比单隐层网络,多隐层网络更容易出现陷入局部极小误差无法摆脱、训练时间急剧增加、训练更加困难和难以达到最佳处理状态等问题。基于以上分析,本试验采用单隐层 BP 神经网络模型。根据 2.2 节,确定模型的输入端为 SCOD、VFAs、甲烷含量、氨氮、当天甲烷产量 5 个输入,选取下一天的产气量作为模型的输出。

2.3.2　隐含层个数与隐含层传递函数的确定

BP 神经网络拓扑结构的关键在于确定隐含层的层数和神经节点的个数,本试验采用公式(1)估计隐含层节点数的大概值。

$$M = \sqrt{p+q} + a \tag{1}$$

式中:M 为隐含层个数;p 为输入层节点数,本试验取 5;q 为输出层节点数,本试验取 1;a 为 1 ~ 10 的常数。

根据公式(1)估算神经网络的隐含层节点数在[3,12]之间,然后采用试凑法来确认隐含层的层数,传递函数的类型选用 S 型传递函数。但是由于 S 型传递函数的种类及隐含层的个数不确定,因此本试验对其进行训练,并选取训练误差最小的指标。为了使训练结果准确度更高,每种组合会训练 10 次,将 10 次训练误差取平均值得到最佳结果(如表 3 所示)。

表 3　模型训练结果

Tab. 3　Model training results

隐含层节点数	隐含层传递函数种类	
	tansig	logsig
3	0.041 9	0.051 0
4	0.033 1	0.051 0
5	0.036 1	0.042 9
6	0.029 9	0.047 1
7	0.029 1	0.035 1
8	0.021 9	0.037 1
9	0.020 4	0.037 1
10	0.028 0	0.029 7
11	0.021 9	0.043 0
12	0.025 5	0.032 7

由表 3 可知,当传递函数为 tansig、隐含层节点数为 9 时,模型的训练误差最小,为 0.020 4。因此,可以确定本试验应用的 BP 神经网络的结构,如图 1 所示。

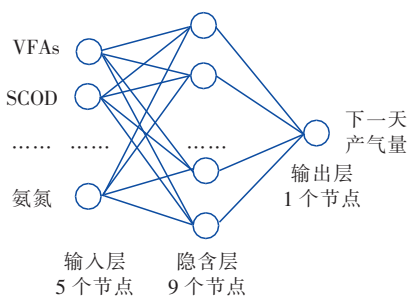


图 1　BP 神经网络拓扑结构

Fig. 1　BP neural network topology

2.3.3　训练函数的确定

本试验采用标准 BP 算法和动量 - 自适应学习速率算法,两种算法对应的训练函数分别为 traingd 和 traingdx。

2.3.4　其他参数的确定

除了前面列举的参数，影响模型精度的指标还包括网络性能目标、学习速率等，本试验利用查阅文献的方法确定这些参数值。例如，由于线性函数具有能够使模型输出任意值的特性，所以通常被用来作为输出层的传递函数。模型参数设定：隐含层到输出层传递函数为线性函数，学习速率为 0.03，网络性能评价指标为均方误差（MSE），动量常数为 0.4，网络性能目标为 0.02，其他参数均为网络自身的默认值。

2.3.5　模型的训练

建模前需要将预处理好的样本集分成训练样本集和测试样本集两部分。两个样本集要同时包含两个反应器的运行数据，得出的测试样本集和训练样本集的数据分别为 10 组和 71 组。模型均方误差的变化如图 2 所示。可知，动量 - 自适应学习速率算法中 BP 神经网络经过 1 890 次训练收敛后，达到的收敛目标误差为 0.02；而标准 BP 算法中神经网络训练的收敛速度较慢，训练时间较长，并陷入了局部最小值中，经过 4 000 次迭代运算后仍不能收敛至目标误差。

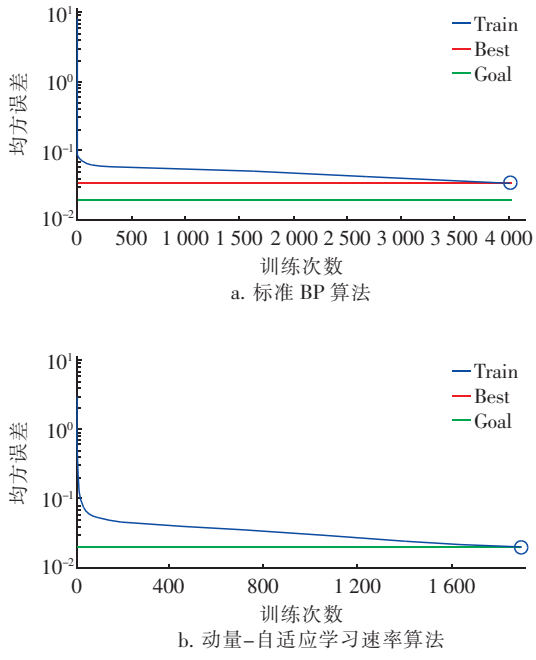


图 2　模型均方误差的变化

Fig. 2　Variation of mean square error of model

两种算法中网络输入值和输出值的线性回归如图 3 所示。可以看出，标准 BP 算法和动量 - 自适应学习速率算法中，BP 神经网络的输出和目标值的线性拟合相关系数分别为 0.87 和 0.91；同时，采用动量 - 自适应学习速率算法获得的拟合方程与直线 $Y=T$ 更贴切，说明该算法具有很强的学习能力，对于模型准确度的提升效果较为明显。

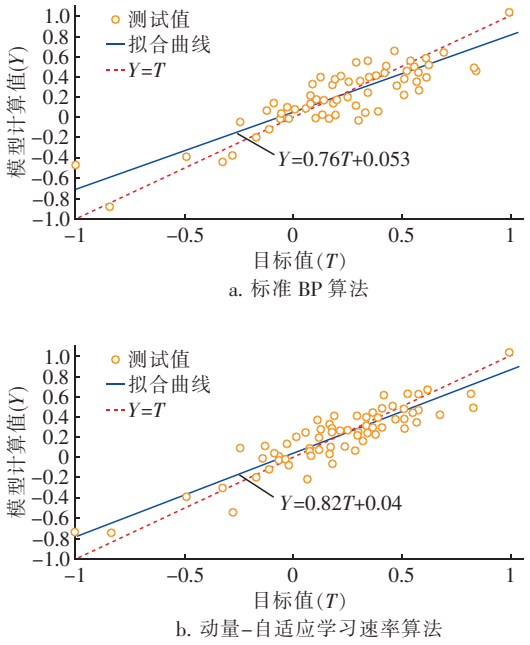


图 3　网络输入和输出值的线性回归

Fig. 3　Linear regression of network input and output values

2.4　模型仿真分析

模型仿真试验数据来自于测试样本集的 10 组数据，将这 10 组数据输入到已经训练完成的模型中进行测试，预测值和实际值的拟合曲线如图 4 所示。

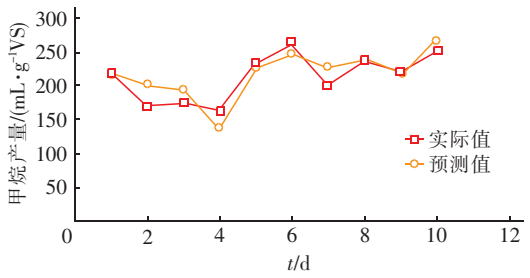


图 4　产气预测值与实际值拟合曲线

Fig. 4　Fitting curve of predicted value and actual value of biogas production

从图 4 可以看出，应用该模型预测的厌氧消化反应器的甲烷产量与实际甲烷产量有较好的相关性，预测值与实际值的变化趋势基本吻合。另外，测试样本集中第 8 组和第 9 组的甲烷产量分别为 238.91 和 223.29 mL/gVS，通过模型预测得到的甲烷产量分别为 239.21 和 218.51 mL/gVS，实际值与

预测值的相对误差仅分别为 0.13% 和 2.14%。预测的准确率较高,说明该模型对于厌氧消化产气量具有一定的预测能力。

3 结论

① 利用相关性分析预处理厌氧反应器的运行指标,确定了模型的输入量和输出量,并利用试凑法确定了单隐层神经网络模型。当输入层到隐含层的传递函数为 tansig、隐含层节点数为 9 时,模型的训练误差最小。

② 动量 - 自适应学习速率算法经过 1 890 次训练收敛后即可以达到目标误差为 0.02,并且输出值与目标值的相关系数为 0.91,更接近直线 $Y = T$,说明动量 - 自适应学习速率算法的拟合性和训练性较好。

③ 对建立好的产气模型进行仿真模拟试验,模拟结果与厌氧反应器实际值相吻合,说明该厌氧产气模型有较好的预测效果。

参考文献:

[1] 宋欢,张光明,王洪臣,等. 污泥与其他基质共消化研究进展[J]. 工业用水与废水,2016,47(4):1 - 6.
Song Huan, Zhang Guangming, Wang Hongchen, *et al*. Research progress of co-digestion of sludge and other substrates[J]. Industrial Water & Wastewater, 2016, 47(4):1 - 6(in Chinese).

[2] 施惠娟. 可视化数据挖掘技术的研究与实现[D]. 上海:华东师范大学,2009.
Shi Huijuan. The Research and Implementation on Visual Data Mining Technology[D]. Shanghai: East China Normal University,2009(in Chinese).

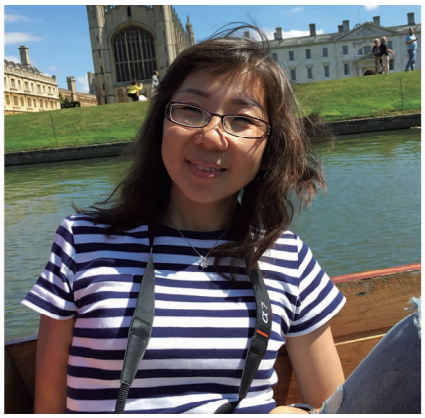
[3] 刘祥明. 水质时间序列数据挖掘及其应用集成研究[D]. 重庆:重庆大学,2011.
Liu Xiangming. Study on Water Quality Time Series Data Mining and Application Integration[D]. Chongqing: Chongqing University,2011(in Chinese).

[4] 范敏. 基于贝叶斯网络的学习与决策方法研究及应用[D]. 重庆:重庆大学,2008.
Fan Min. Research and Application on Learning & Decision Methods Based on Bayesian Network[D]. Chongqing: Chongqing University,2008(in Chinese).

[5] 张文阳,张良均,李娜,等. 多元回归和 BP 人工神经网络在预测混合厌氧消化产气量过程中的应用比较[J]. 环境工程学报,2013,7(2):747 - 752.
Zhang Wenyang, Zhang Liangjun, Li Na, *et al*. Comparing multiple regression and BP artificial nerve net model used on prediction of anaerobic co-digestion gas-producing process[J]. Chinese Journal of Environmental Engineering, 2013,7(2):747 - 752(in Chinese).

[6] 奇敏. 有机垃圾厌氧消化产气规律及模型研究[D]. 武汉:华中科技大学,2009.
Qi Min. The Research on the Regularity and Model of Anaerobic Digestion of the Organic Waste[D]. Wuhan: Huazhong University of Science & Technology, 2009(in Chinese).

[7] 赖夏颉. 基于数据挖掘技术的厌氧消化系统模拟研究[D]. 成都:西南交通大学,2014.
Lai Xiajie. Anaerobic Digestion System Simulation Research Based on Data Mining Technology[D]. Chengdu: Southwest Jiaotong University,2014(in Chinese).



作者简介:温沁雪(1977 -),女,黑龙江哈尔滨人,博士,副教授,主要从事废物资源化理论与技术研究。

E - mail:wqxshelly@263.net

收稿日期:2018 - 08 - 22