The Eurasia Proceedings of Science, Technology, Engineering & Mathematics (EPSTEM)
ISSN: 2602-3199

The Eurasia Proceedings of Science, Technology, Engineering & Mathematics (EPSTEM), 2018

Volume 4, Pages 212-217

IConTES 2018: International Conference on Technology, Engineering and Science

Prediction Modeling of Biogas Production with Classification and Regression Tree at Wastewater Treatment Plants

Halil AKBAS
Süleyman Demirel University

Gultekin OZDEMIR
Süleyman Demirel University

Abstract: Predicting biogas production is important for energy management in wastewater treatment plants (WWTPs). Biogas production quantity depends on its production system variables, such as, influent flow rate, process temperature, alkalinity, volatile fatty acid, sludge retention time, total suspended solid, etc. WWTPs keep the records of wastewater treatment process values with supervisory control and data acquisition (SCADA) system on a regular basis. The relationship between the biogas production and its production system variables, which are measured continuously with SCADA system, can be identified with classification and regression tree (CART) algorithm by using the existing data. In this paper, CART approach is presented for the prediction of biogas production at WWTPs. Standard CART algorithm is used to select split predictor. Curvature and interaction tests are also applied in the model to search for reducing split predictor selection bias and improving the detection of important interactions among each predictor and response and among each pair of predictors and response in turn.

Keywords: Prediction, Classification and regression tree, Biogas production, Wastewater treatment plant

Introduction

Renewable energy has come into prominence in the world in recent years. Anaerobic sludge digestion systems are one of the important technologies, which have been used at WWTPs in Turkey. Anaerobic digesters produce biogas with the gasification of sludge and obtained biogas is used as the fuel of cogeneration unit. Cogeneration system converts biogas to heat and electrical energy. The cost of electricity is one of the highest item of the expenditure budget of WWTPs, if electricity is purchased from service providers. Therefore, cogeneration systems at WWTPs are connected to grid to cover economically the electrical energy need of WWTPs.

There are some studies in literature related with prediction models of wastewater treatment, anaerobic digestion, biogas production and biogas quality at WWTPs. Artificial intelligence methods, e.g., artificial neural networks (ANNs) and adaptive neural fuzzy inference system (ANFIS) have been successfully used for prediction modeling in this field (Tay and Zhang, 1999; Holubar et al., 2002; Strik et al., 2005; Cakmakci, 2007; Kusiak and Wei, 2011, Zhang et al., 2012). With these articles, artificial intelligence algorithms accurately converged the predicted outputs to observed outputs. It was presented that ANFIS and ANN were potential prediction models to simulate and control complex wastewater treatment systems at WWTPs.

Machine learning aims to teach computers to do what comes naturally to humans. By using computational methods, they learn information directly from data and do not rely on a predetermined equation as a model. Machine learning algorithms` prediction accuracy intends to improve, while the number of samples available for learning increases. SCADA technologies enable the continuous collection of data. The collected data contains process information, which can be used with machine learning algorithms. Useful patterns and models are identified by running machine learning algorithms on the basis of statistics and computational intelligence (Witten et al., 2011). The applications of these approaches have been presented with business, manufacturing,

- Selection and peer-review under responsibility of the Organizing Committee of the Conference

science, and engineering publications in literature (Takada et al., 1995; Wang et al., 1995; Shah et al., 2006; Kusiak et al., 2009 & 2010; Kusiak and Smith, 2007).

Machine learning uses two types of techniques, which are supervised and unsupervised learning. Supervised learning trains a model on known input and output data to build a model that generates reasonable predictions based on evidence for the response to new data. Supervised learning uses classification and regression techniques to develop predictive models. Classification techniques predict categorical responses, for example, whether a tumor is malignant or benign. On the other hand, regression techniques predict continuous responses, for instance, changes in temperature or fluctuations in power demand. In the course of prediction model development, finding the most suitable algorithm is partly based on trial and error (Witten et al., 2011).

The aim of this article is to establish the prediction modeling of biogas production system at WWTPs by using CARTs. Prediction modeling of biogas production system is important for energy efficiency at WWTPs. As presented in Figure 1, biogas production system at WWTPs typically has anaerobic digesters, gas storage system, flare and desulphurization units to produce biogas and to power cogeneration unit for heat and electrical energy production. Sludge digestion process provides biogas production at WWTPs. Prediction models of anaerobic sludge digesters with system variables having complicated and non-linear relationships can be set by using supervised machine learning techniques. This article dwells on the prediction of biogas production at WWTPs and proposes a CART prediction model. The proposed prediction model is established by considering system data of a WWTP in Turkey. With this article, the goal of extended use of machine learning techniques to establish prediction models of renewable energy systems in Turkey is pursued.

Methodology

Data Description

The daily data of the biogas production system at the WWTP in Turkey is used. Firstly, the hourly data is arranged. Then, the average of hourly data is taken to apply in the prediction model. The data of different departments contributing to biogas production at the WWTP is put together to form an aggregated data set. Intrusive data in the periods of system failure, machine breakdown and scheduled maintenance are identified as outliers. Processed data set includes 776 suitable data points. Data set is divided into training and testing data sets with 50% share for each. Training data is used to train and develop a prediction model with CART, while test data is used for model validation.

The list of input variables, which are temperature (T), pH, sludge loading rate (SLR), total suspended solid (TSS) and total volatile solid (TVS) and output, which is biogas production (BP), with their minimum and maximum values are presented in Table 1.

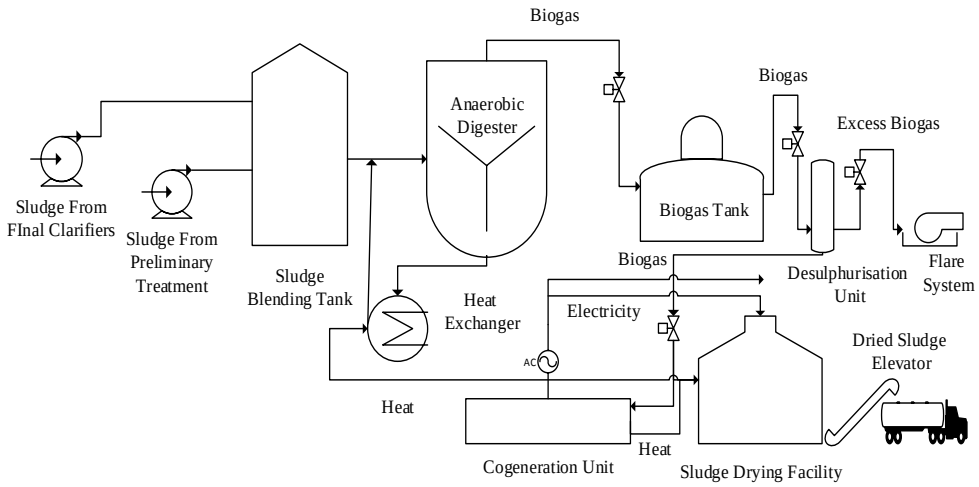

Figure 1. An example of biogas production system at WWTPs (Akbas et al., 2015)

Table 1. List of inputs and output

Input Variable	Unit	Min.	Max.
T	°C	31.02	37.18
pH	-	6.62	7.65
SLR	m3/day	129	592
TSS	mg/lt	19500	33555
TVS	mg/lt	13075	23413
Output			
Biogas Production	m3/day	1167	4096

Prediction Model

CART is a prediction method, which predicts responses to data by following the decisions in the tree from the root node to down to leaf node. Leaf nodes contain the response. Classification trees typically give responses that are not nominal, such as, true or false, while regression trees provide numeric responses (Breiman et al., 1984).

A regression tree is either a piecewise constant or piecewise linear estimate of a regression function. The data is recursively partitioned to construct regression trees. In this article, the inputs of prediction model are continuous variables within a constant interval, while model output is a numeric response. For this reason, regression trees are used for the prediction of numeric output.

Loh (2002) stated that the first implementation of the idea of CART was proposed as the AID algorithm. With the AID algorithm, a piecewise constant estimate is calculated by searching over all axis-orthogonal partitions. The binary partition with minimum total sum of the squared error (SSE) is selected. A pre-determined decrease rate of SSE is used to decide when to stop splitting. SSE can be calculated with equation (1) as given below.

$$\mathrm{SSE} = \sum_{i=1}^{n}(y_i - \bar{y})^2 \tag{1}$$

There are two main weaknesses of AID algorithm, First one is that choosing the best decrease rate is experimental and it can lead to over or under fitting. The other weakness is that greedy search approach can cause bias in variable selection (Loh, 2002). However, CART algorithm uses backward elimination approach to prevent from over or under fitting while creating decision trees. It grows large trees, prunes away some branches and uses test data or cross-validation to calculate the total SSE. CART algorithm uses greedy search of AID algorithm for variable selection and still incorporates the problem of selection bias (Loh, 2002).

When the prediction model has continuous predictors having many levels, standard CART algorithm can be useful. Standard CART approach aims to choose the split predictor with maximal split criterion gain over all possible splits of all predictors. Loh (2002) presented an algorithm called generalized, unbiased interaction detection and estimation (GUIDE) with useful features. If there are categorical predictors with fewer levels or interaction between pairs of predictors and the response are considered, curvature test and interaction detection can be used as split predictor selection techniques. Curvature test chooses the split predictor with minimum p-value of chi-square tests of independence between each predictor and the response, while interaction-curvature selects the split predictor with minimizing the p-value of chi-square tests of independence between both each predictor and response and each pair of predictors and response (Loh, 2002).

Prediction Accuracy

Prediction accuracy for CART model is tested by using statistical indicators. Root mean square error (RMSE) and regression coefficient are calculated for observed outputs and predicted outputs by using equations in (2) and (3).

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left|\hat{y}_i - y_i\right|^2} \tag{2}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - y_m)^2} \tag{3}$$

where;

$\hat{y}_i$: predicted output value

y_m : the mean of observed output values
n : the number of total instances in the data set

Results and Discussion

Prediction Results

The proposed prediction models are implemented in Matlab R2017b version. The experiments have been conducted using a notebook with 2.40 GHz Intel (R) Core™ processor and 8.00 GB RAM.

The data set includes training and testing data of observed values of BP (m3/day). BP per day has been placed into diagram along with the corresponding values predicted by CART algorithm with CART, curvature and interaction-curvature split predictor selection methods as presented in Figures 1, 2 and 3 in order.

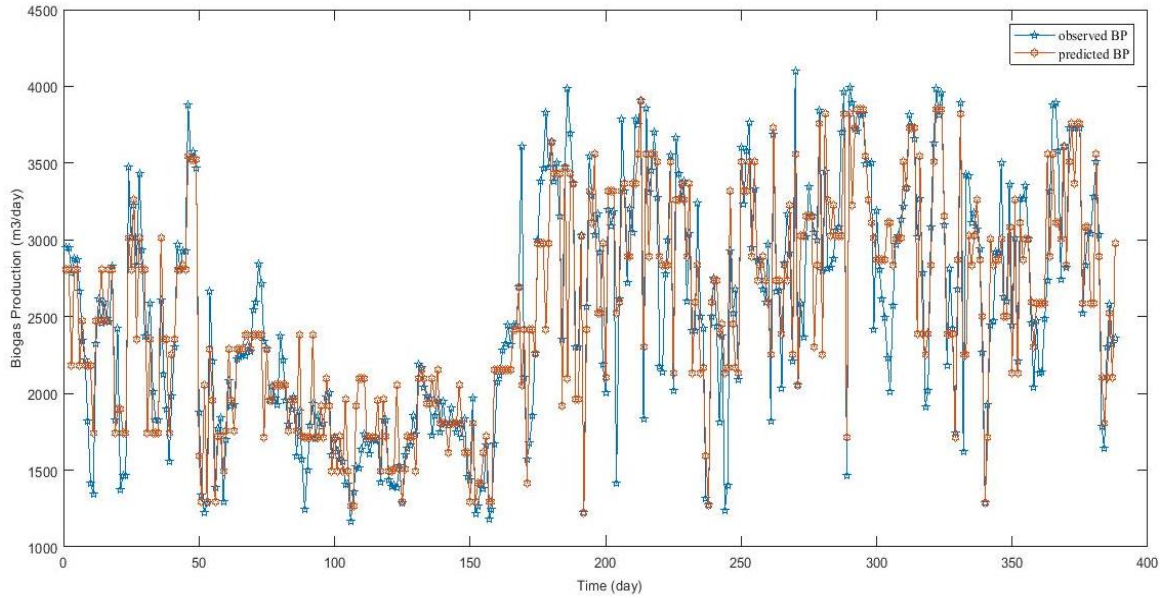

Figure 1. CART model observed and predicted biogas production

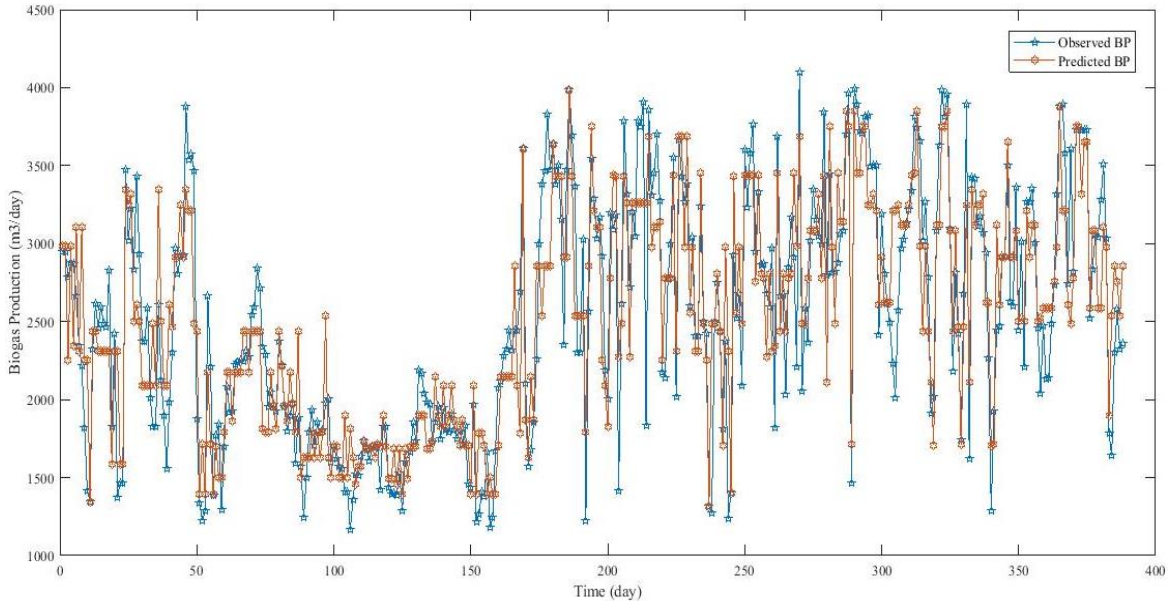

Figure 2. CART model with curvature test observed and predicted biogas production

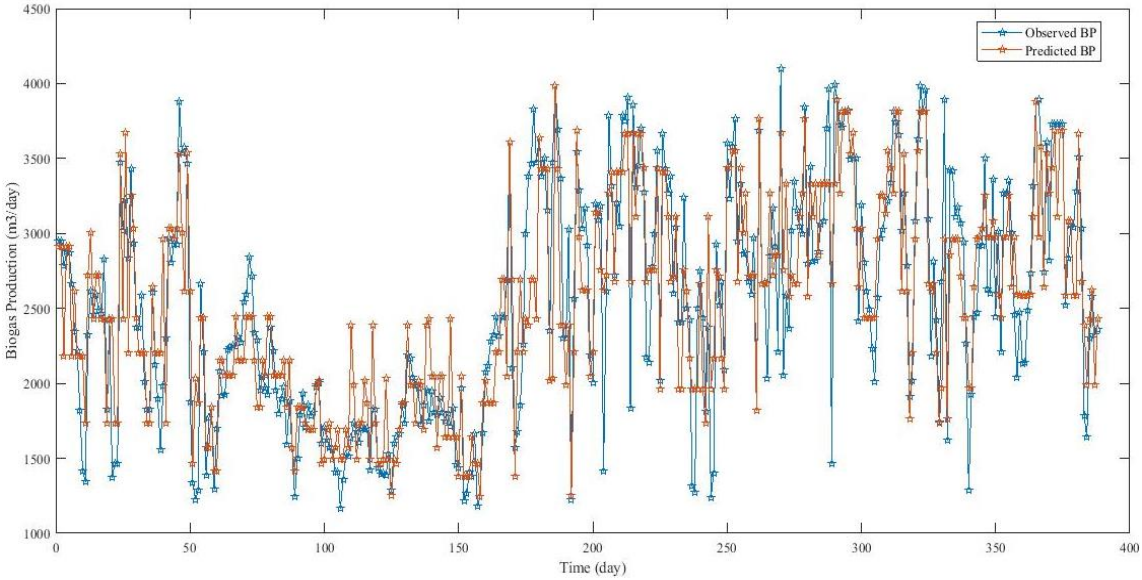

Figure 3. CART model with interaction-curvature test observed and predicted biogas production

All of the figures are drawn by using only test data with 388 data points. The blue lines indicate the observed BP values and the red lines present the predicted BP values. Figures 1, 2 and 3 show that some of the observed values are very closely predicted. However, it is seen that a number of predicted values are either larger or smaller than observed values.

Prediction Accuracy Results

Regarding to the prediction accuracy results of the CART model, three CART prediction models are established by using CART algorithm, curvature test and interaction-curvature test as split selection algorithm of regression trees. The obtained results of prediction accuracy are presented in Table 2.

Table 2. Prediction accuracy results of CART algorithm

Split predictor selection technique	Regression coefficient	RMSE	Elapsed time (sec.)
CART	0.80	13.14	2.732318
Curvature test	0.78	13.77	1.251434
Interaction-curvature test	0.76	14.89	1.547569

CART algorithm has the best regression coefficient, which is 0.80, between the observed and predicted BP, when the CART split selection technique is used. However, elapsed time of prediction with CART split selection is longer than both curvature and interaction-curvature tests. The shortest elapsed time is obtained with curvature split selection technique. On the other hand, CART algorithm with curvature and interaction-curvature split selection techniques also predicted the observed BP values with 0.78 and 0.76 regression coefficients in turn closely to the prediction with CART split selection technique. The calculated values of RMSE are quite small and close to each other for all prediction models as given in Table 2.

Conclusion

Three CART models are used in the design of intelligent systems to predict the biogas production system at a WWTP in Turkey. CART prediction models are prepared with split variable selection techniques, which are CART, curvature test and interaction-curvature test, for calculating the best predicted BP response values. It is seen that CART prediction model with standard CART approach to select split predictors with maximum split criterion gain over all possible splits of predictors provides the most accurate prediction with the regression coefficient value of 0,80. However, the methodology of analyzing the interactions among pairs of predictors and response of curvature and interaction detection tests can also be used as split predictor selection techniques of

CART prediction model to provide the control of the biogas production system consistently with the regression coefficient values of 0.78 and 0.76 in order.

References

Akbas, H., Bilgen, B, & Turhan, A.M. (2015). An integrated prediction and optimization model of biogas production system at a wastewater treatment facility. Bioresource Technology, 196, 566-576.

Breiman, L., Friedman, J.H., Olshen, R.A., & Stone, C.J. (Eds.). (1984). *Classification and regression trees*. Florida, FL: Chapman & Hall/CRC Press.

Cakmakci, M. (2007). Adaptive neuro-fuzzy modelling of anaerobic digestion of primary sedimentation sludge. Bioprocess Biosystem Engineering, 30, 349-357.

Holubar, P., Zani, L., Hager, M., Froschl, W., Zorana, R., & Braun, R. (2002). Advanced controlling of anaerobic digestion by means of hierarchical neural networks. Water Research, 36, 2582-2588.

Kusiak, A, & Wei, X. (2011). Prediction of methane production in wastewater treatment facility: A data-mining approach. Annals of Operations Research, 216, 71-81.

Kusiak, A., & Smith, M. (2007). Data mining in design of products and production systems. Annual Reviews in Control, 31, 147-156.

Kusiak, A., Zheng, H. Y., & Song, Z. (2009). Wind farm power prediction: a data-mining approach. Wind Energy, 12, 275–293.

Kusiak, A., Li, M. Y., & Tang, F. (2010). Modeling and optimization of HVAC energy consumption. Applied Energy, 87, 3092–3102.

Loh, W. Y. (2002). Regression trees with unbiased variable selection and interaction detection. Statistica Sinica, 12, 361-386.

Shah, S., Kusiak, A., & O'Donnell, M. (2006). Patient-recognition data-mining model for BCG-plus interferon immunotherapy bladder cancer treatment. Computers in Biology and Medicine, 36, 634–655.

Strik, D., Domnanovich, A.M., Zani, L., Braun, R., & Holubar, P. (2005). Prediction of trace compounds in biogas from anaerobic digestion using the MATLAB neural network toolbox. Environmental Modelling & Software, 20, 803-810.

Takada, T., Sanou, K., & Fukumara, S. (1995). A neural network system for solving an assortment problem in the steel industry. Annals of Operations Research, 57, 265–281.

Tay, J.H., & Zhang, X. (1999). Neural fuzzy modeling of anaerobic biological wastewater treatment systems. Journal of Environmental Engineering, 125, 1149-1159.

Wang, Q., Sun, X., Golden, B. L., & Jia, J. (1995). Using artificial neural networks to solve the problem. Annals of Operations Research, 61, 111–120.

Witten, I.H., Frank, E., & Hall, M.A. (3rd Eds.). (2011). *Data mining practical machine learning tools and techniques*. Burlington, MA: Morgan Kaufmann Publishers.

Zhang, Z., Zeng, Y., & Kusiak A., (2012). Minimizing pump energy in a wastewater processing plant. Energy, 47, 505-514.

Author Information

Halil Akbas
Süleyman Demirel University, Graduate School of Natural and Applied Sciences, Department of Industrial Engineering, 32260, Isparta, Turkey
Tel.: +90 (246) 211 38 47
Fax.: +90 (246) 237 10 19
Contact E-mail: *akbas.halil@yahoo.com*

Gultekin Ozdemir
Süleyman Demirel University, Faculty of Engineering, Department of Industrial Engineering, 32260, Isparta, Turkey.
Tel.: +90 (246) 211 82 45
Fax.: +90 (246) 237 08 59