Advanced Engineering Informatics 43 (2020) 101030

Contents lists available at ScienceDirect

Advanced Engineering Informatics

journal homepage: www.elsevier.com/locate/aei

Full length article

Ensemble data mining modeling in corrosion of concrete sewer: A comparative study of network-based (MLPNN & RBFNN) and tree-based (RF, CHAID, & CART) models

Mohammad Zounemat-Kermani[a,*], Dietmar Stephan[b], Matthias Barjenbruch[c], Reinhard Hinkelmann[d]

[a] Department of Water Engineering, Shahid Bahonar University of Kerman, Kerman, Iran
[b] Chair of Building Materials and Construction Chemistry, Institute of Civil Engineering, Technische Universität Berlin, Berlin, Germany
[c] Chair of Urban Water Management, Institute of Civil Engineering, Technische Universität Berlin, Berlin, Germany
[d] Chair of Water Resources Management and Modeling of Hydrosystems, Institute of Civil Engineering, Technische Universität Berlin, Berlin, Germany

ARTICLE INFO

Keywords:
Concrete corrosion
Machine learning
Soft computing
Sewer systems
Artificial intelligence

ABSTRACT

This research aims to evaluate ensemble learning (bagging, boosting, and modified bagging) potential in predicting microbially induced concrete corrosion in sewer systems from the data mining (DM) perspective. Particular focus is laid on ensemble techniques for network-based DM methods, including multi-layer perceptron neural network (MLPNN) and radial basis function neural network (RBFNN) as well as tree-based DM methods, such as chi-square automatic interaction detector (CHAID), classification and regression tree (CART), and random forests (RF). Hence, an interdisciplinary approach is presented by combining findings from material sciences and hydrochemistry as well as data mining analyses to predict concrete corrosion. The effective factors on concrete corrosion such as time, gas temperature, gas-phase H_2S concentration, relative humidity, *pH*, and exposure phase are considered as the models' inputs. All 433 datasets are randomly selected to construct an individual model and twenty component models of boosting, bagging, and modified bagging based on training, validating, and testing for each DM base learners. Considering some model performance indices, (e.g., Root mean square error, *RMSE*; mean absolute percentage error, *MAPE*; correlation coefficient, *r*) the best ensemble predictive models are selected. The results obtained indicate that the prediction ability of the random forests DM model is superior to the other ensemble learners, followed by the ensemble Bag-CHAID method. On average, the ensemble tree-based models acted better than the ensemble network-based models; nevertheless, it was also found that taking the advantages of ensemble learning would enhance the general performance of individual DM models by more than 10%.

1. Introduction

1.1. Background

The maintenance and rehabilitation of sewer networks, which are the main infrastructures for sewage and wastewater transport in cities, is not only a complex procedure but also a costly process. In this respect, the microbially induced concrete corrosion (MICC) of sewer collection and conveyance systems is recognized as one of the main factors for degradation of the sewer infrastructure and reducing the lifetime of concrete structures such as pipes and manholes. MICC in sewers resulted from a combination of physicochemical and biological processes under the acidic condition of the internal concrete surfaces. In other words, the favorable corrosive conditions for anaerobic sulfate-reducing bacteria occurred due to high levels of relative humidity, high concentration of CO_2 and H_2S, and lowered surface pH [24,14,10].

Having the ability to predict the rate of MICC accurately, would lead to better estimate the sewer service life, planning sewer maintenance, and rehabilitation process. Nevertheless, the reliable prediction of MICC (as a function of time) has been considered difficult due to the complexity of the factors involved, for instance, sewer atmosphere temperature, relative humidity, surface pH, number of abiotic and biotic processes, concrete porosity, concrete type, amount of sulfur, carbon, nitrogen and mineral salts in the environment [31]. Ideally, a predictive model should consider all of the effective parameters on MICC in sewers; however, some of the above parameters cannot, or

* Corresponding author.
E-mail address: zounemat@uk.ac.ir (M. Zounemat-Kermani).

https://doi.org/10.1016/j.aei.2019.101030
Received 28 March 2019; Received in revised form 25 November 2019; Accepted 13 December 2019

might not, be measured in laboratory tests and in-situ measurements. Nonetheless, simple regression and empirical models are not potential candidates for considering all available measured factors for establishing a proper predictive model. Indeed, data mining approaches can be considered as alternative appraoches for coping with this problem.

Data mining approaches combine database technology, machine learning (ML) and statistical analysis for the sake of extracting hidden patterns, information and relationships from large databases. Generally, there are two data mining strategies for modeling and predicting processes: the standard approaches which are constructed on one individual model, and the more recent ensemble learning algorithms (e.g. bagging, boosting, random forests), which are established based on several component models of the data mining base learners [2]. Ensemble learning is an effective technique so that previous studies concur that the ensemble learning algorithms have higher accuracy than the accuracy of a standard individual DM model [27]. Hence, the use of advanced ensemble data mining techniques, like boosting and bagging, can help knowledge discovery in modeling and predicting complex phenomena, such as MICC.

1.2. Literature review

Several researchers have developed a traditional approach for constructing data mining models for modeling and predicting the MICC initiation and corrosive rate. Alani and Faramarzi [1] proposed evolutionary polynomial regression (EPR) technique to predict the degradation of concrete subject to sulfuric acid attack. The results showed that the EPR model can successfully predict mass losses of concrete specimens exposed to sulfuric acid. Jiang et al. [15] presented an artificial neural network (ANN) for predicting the initiation time and the corrosion rate in sewers. The ANN trained model estimated the corrosion initiation time and corrosion rates very close to those measured in situ. Liu et al. [19] proposed a hybrid Gaussian processes regression (HGPR) model to simulate the evolution of the concrete corrosion rate. The temperature and H_2S data were considered as the main element for constructing the simulative model. The HGPR model presented better results in terms of statistical measures compared with the multiple linear regression model (MLR) and radial basis function neural network. ANNs were employed by Hendi et al. [12] to predict the mass and volume-loss under different conditions for the durable performance of concrete subjected to H_2SO_4 attacks of sewage systems. Based on the ANN analysis, higher microsilica and glass powder contents and concretes enhance the performance of concrete in an H_2SO_4 acid medium. Li et al. [17] used three different data-driven models (MLR, ANN, and adaptive neuro-fuzzy inference system (ANFIS)) for predicting the corrosion initiation time and corrosion rate in sewers. The results of the study proved that the ANN and ANFIS models performed better than the MLR model for the corrosion prediction.

Though there are not any researches reported of the application of ensemble data mining models for concrete corrosion, some studies have described the use of ensemble ML models in modeling concrete characteristics, such as compressive strength. Erdal et al. [9] incorporated bagging and gradient boosting methods in building ANN ensembles for high-performance concrete (HPC) compressive strength forecasting. The results of the study showed that ensemble models are promising techniques on HPC compressive strength forecasting, and they were superior to the conventional applied ANN model. It was also concluded that both the bagging and gradient boosting methods gave nearly the same results in the forecasting process. Chou et al. [7] provided a comparative study using individual and ensemble (e.g., bagging) machine learning techniques to predict the compressive strength of HPC. Individual and ensemble learning classifiers were constructed from four different base learners, including MLPNN, support vector machine (SVM), classification and regression tree (CART), and linear regression (LR). The comparison results showed that ensemble learning techniques were better than learning techniques used individually to predict HPC compressive strength. Aydogmus et al. [3] investigated the potential usage of bagging ensemble approach and four base artificial intelligence models for modeling concrete slump flow. It was reported that the bagging ensemble models were superior to the individual base learner models and reduced the prediction error of proposed predictive models.

Recently, there has been a lot of research in enhancing the performance of ML models by generating classifiers and aggregate their outcomes using so-called ensemble-learning methods. Indeed, recent classification and prediction modeling researches have proven that ensemble techniques have gained popularity and become more common in practice as they usually perform better than individual base learners do [11,25]. Pham et al. [22] assessed and compared the performance of several ML ensemble technique including Bagging, adaptive boosting (AdaBoost), Dagging, Rotation Forest, MultiBoost, and Random SubSpace for the evaluation of landslide susceptibility. They used MLPNN as the base learner model for the mentioned ensemble learning methods. They concluded that the performance of individual MLPNN was enhanced by the use of ensemble learning. Besides, they pointed out that the MultiBoost technique gave the best results compared to the other ensemble methods. Hamori et al. [11] predicted the payment data using three types of ensemble learning methods (bagging, random forests, and boosting) based on an artificial neural network model. The results obtained showed that the boosting technique was superior to other applied learning methods.

Papadopoulos et al. [21] evaluated the performance of three ensemble learning techniques such as random forests, gradient boosted regression trees, and extremely randomized trees in predicting the heating and cooling loads of buildings. Results showed that the ensemble techniques outperformed the traditional individual machine learning models. To be more precise, the gradient boosting technique gave the best performance in the prediction process. Saeed et al. [25] developed an ensemble bagged tree (EBT) algorithm for detection of power distribution companies. The accuracy of the EBT algorithm for non-technical losses detection was found to be higher than individual machine learning models such as decision trees and support vector machine.

1.3. Rationale, contribution, and objectives

The majority of sewer corrosion studies are limited to the use of experimental and laboratory tests, field studies, and numerical simulations. On the one hand, summarizing the findings of the restricted literature review on using data mining (DM) approaches for simulating concrete corrosion in sewers proves the eligibility of these methods for modeling the MICC process. On the other hand, employing ensemble techniques can improve the performance of machine learning (ML) methods.

Hence, the present research follows the main objective of the use of DM approaches in MICC prediction in sewer systems. The efficiency of individual and ensemble boosting and bagging ML models (including Multi-layer perceptron neural network, MLPNN; radial basis function neural network, RBFNN; chi-square automatic interaction detector, CHAID; and classification and regression tree, CART) as well as modified bagging model (random forests, RF) in predicting time-dependent mass loses of MICC in sewers will be evaluated. For meeting this end, in addition to the time factor, several environmental key factors will be utilized as independent variables such as relative humidity, gas temperature, surface concrete *pH*, H_2S concentration and gas/submerged status of the exposed concrete.

The contribution and novelty of this study are twofold. A primary contribution of this study is to challenge the ability of network-based (MLPNN & RBFNN) and tree-based (CHAID, CART, RF) models taking into account three types of ensemble learning methods including the i) bagging and ii) boosting and iii) modified bagging techniques. Despite the fact that there have been several reports on using different types of