Automation in Construction 44 (2014) 84–91

Contents lists available at ScienceDirect

Automation in Construction

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

Evaluation of artificial intelligence tool performance and uncertainty for predicting sewer structural condition

Vitor Sousa *, José P. Matos, Natércia Matias

DECivil, Instituto Superior Técnico, Universidade de Lisboa, Av. Rovisco Pais, 1049-001 Lisbon, Portugal

ARTICLE INFO

Article history:
Received 26 November 2013
Received and Revised form 26 March 2014
Accepted 5 April 2014
Available online 26 April 2014

Keywords:
Artificial neural networks
Optimization
Sewer structural condition
Support vector machines

ABSTRACT

The implementation of a risk-informed asset management system by a wastewater infrastructure utility requires information regarding the probability and the consequences of component failures. This paper focuses on the former, evaluating the performance of artificial intelligence tools, namely artificial neural networks (ANNs) and support vector machines (SVMs), in predicting the structural condition of sewers. The performance of these tools is compared with that of logistic regression on the case study of the wastewater infrastructures of SANEST — *Sistema de Saneamento da Costa do Estoril* (Costa do Estoril Wastewater System). The uncertainty associated to ANNs and SVMs is quantified and the results of a trial and error approach and the use of optimization algorithms to develop SVMs are compared. The results highlight the need to account for both the performance and the uncertainty in the process of choosing the best model to estimate the sewer condition, since the ANNs present the highest average performance (78.5% correct predictions in the test sample) but also the highest dispersion of performance results (73% to 81% correct predictions in the test sample), whereas the SVMs have lower average performance (71.1% without optimization and 72.6% with the parameters optimized using the Covariance Matrix Adaptation Evolution Strategy) but little variability.

1. Introduction

During the last decades there has been a trend to develop and implement formal asset management systems for wastewater infrastructures. These asset management systems have been gradually evolving from reactive to proactive stances and their scope has broadened significantly to the point of being considered the central element in the technical management of water and wastewater infrastructures [1–6].

One of the first proactive-based asset management systems was developed by the Water Research Centre, in which the defects observed during Closed-Circuit Television (CCTV) inspections were rated in order to obtain a classification for the sewer condition. Originally, the approach was used only to manage the critical sewers, that is, managing proactively the sewers that entail very high economic consequences in case of failure, and reactively the remaining [7]. However, due to the growing awareness of the non-economic dimension of sewer failures, the application of this approach was expanded to the non-critical assets [8]. This approach has been implemented worldwide, with adjustments introduced by national institutions and local municipalities [9,10].

More complex and comprehensive models were also developed with the purpose of optimizing decisions and prioritizing interventions, by taking into account hydraulic, environmental, social and economic constrains (MARESS — [11]; RERAUVIS — [12]; CARE-S — [13]). Additionally, there is a growing demand for conducting periodical sewer inspections in order to comply with legal requirements (e.g., in Germany, most States require the inspection of the total sewer network once in ten years). This has led to the development of models for assisting decisions regarding which sewers are to be inspected (AQUA-WertMin — [14]; SCRAPS — [15–17]).

If there is the need to support rehabilitation or inspection decisions, the models developed to predict interventions in sewer systems should include a module for estimating the evolution of the sewer condition [18]. Most of such models either require information that is not always available and is usually not easily obtained (e.g., soil aggressiveness) or are based on statistical analysis. The traditional statistical models require the previous knowledge of the function/structure that better represents the effect of the different sewer characteristics (e.g., material, diameter, age) on its performance. This is a major drawback because the effect of the interactions between different sewer characteristics and how they relate between them and with the sewer performance is not known neither easy to determine (e.g. diameter and age interact as sum, a product, a power, a logarithm or any other mathematical formulation). Consequently, in most cases, the statistical models consider only one (usually the age) or two of these characteristics (usually the age in combination with one of the others). Artificial intelligence tools are an alternative that can be used in classification and pattern identification problems such as this (e.g. [19]). The present paper discusses the use of artificial neural networks (ANNs) and support vector machines (SVMs) to estimate the condition of sewers, being the results compared with those of

---

* Corresponding author. Tel.: +351 218418381.
E-mail address: vitor.sousa@tecnico.ulisboa.pt (V. Sousa).

http://dx.doi.org/10.1016/j.autcon.2014.04.004
0926-5805/

logistic regression. The models were fitted using data from the periodic CCTV inspection program that has been implemented in the SANEST — *Sistema de Saneamento da Costa do Estoril* (Costa do Estoril Wastewater System) since 2005.

2. Sewer condition modeling

2.1. Approaches

Depending on the type of output provided, the models for determining sewer conditions can be classified as deterministic or stochastic [20,21]. The former provide estimates in an absolute and exact format while the latter comprise some form of uncertainty/variability quantification associated with the estimate. The deterministic models are the most common, but in a risk-informed context it is important to evaluate the uncertainty of the estimates in order to determine the best options. The models for predicting sewer conditions can be further classified into empirical or mechanistic [20,21]. The empirical models use statistical tools and methods to obtain relations between known variables and the sewer condition based on historical records. These models entail an implicit assumption that the pattern of deterioration will remain the same in the future. The mechanistic models seek to represent the physical, chemical and/or biological phenomena that take place within the sewers and are relevant to explain their condition. These models require information and data that is not generally available or cannot be easily obtained. There is another class of models (expert-based models) that rely on expert opinion to define the relation between the inputs and the outputs. Historically, these models (e.g. MOSIMO — [22,23]; SCRAPS — [15–17]) have been less explored and are seldom put into practice, although for situations of information scarcity they may be the only viable option for estimating the sewer condition.

Empirical models are the most widely studied. These models, also called statistical models, are built from statistical analyses of operation and maintenance failure records (e.g., clogging; collapse) or condition classification based on inspection data (e.g., operational or structural condition classification using a rating protocol). Two main categories of empirical models can be identified [19,24]: i) function-based models; and ii) data-based models. Both model categories rely on fitting observed data. However, for function-based models the mathematical expressions relating the inputs with the outputs are pre-defined at the outset. In this case, the fitting operation seeks to determine the coefficients of the functions that minimize the error between the observed and the estimated outputs. In the data-based models there is no pre-defined expression relating the inputs with the outputs such that the fitting operation simultaneously adjusts the relation between the inputs and the outputs and the relative weight of each input. Table 1 resumes the main classes and types of function-based and data-based models that have been used for estimating sewer condition.

The references presented in Table 1 are specific to sewer systems. There are also similar studies on water supply networks that use alternative empirical models which could be adapted for sewer systems (e.g., [55–58]).

The present paper focuses on the application of ANNs and SVMs, which are machine learning techniques. These techniques have the ability to learn the patterns of the underlying process from past data and generalize the relationships between input and output data, being able to predict or estimate an output given a new set of input variables from the vicinity of the training domain. Some brief details on these techniques are provided next, along with a review on logistic regression.

2.2. Logistic regression (LR)

The logistic regression (LR) is a type of generalized linear model that extends the linear regression by linking the range of real numbers to the 0–1 range, allowing approaching situations where the response variable is qualitative and takes on only two possible values [59]. LR assumes the response to be a Bernoulli random variable and provides a prediction of the chance that the response will assume one of the categorical response levels [60]. Considering that the value 1 represents the event of interest, the relation between the probability of it happening ($P[y_k = 1]$) and the predictors (x_i) is given by a logistic model:

$$P[y_k = 1] = p_k = \frac{1}{1 - e^{-\left(\beta_{ok} + \sum_{i=1}^{n} \beta_{ik} x_i\right)}} \tag{1}$$

where p_k is the probability that the kth case experiences the event of interest; β_{ik} is the value of the ith regression coefficient of the kth case; x_i is the ith predictor; and n is the number of predictors. If $p_k \geq 0.5$ the case falls into class 1, otherwise it falls into class 0. The model assumes that the predictors are not highly correlated since, as in the linear regression, this can cause problems with the estimation of the coefficients [61]. Nonetheless, LR is regarded to be robust even when the assumptions are not fully met [62]. Usually, the coefficients β_{ik} are obtained using the maximum likelihood estimates [63].

The logistic function provides a mean for mapping from the predictor domain onto the [0, 1] interval [64]. Other commonly used function is the normal probability distribution, resulting in the so called probit regression model. Another alternative, the multinomial logistic regression, expands the LR for response variables with $m > 2$ classes. In this case, there will be $m - 1$ complementary link functions.

2.3. Artificial neural networks (ANNs)

Since Warren S. McCulloch and Walter H. Pitts proposed the first artificial model for a biological neuron of human brain in 1943, numerous methods for building neuro-inspired computational models have been proposed and investigated [65]. ANNs are defined not only by their use of artificial neurons, but also by the network structure connecting them. In addition, there are other important features to be

Table 1
Empirical models used for estimating the condition of sewers.

Category	Class	Type	References
Function-based	Deterministic	Linear regression	[25–27]
		Non-linear regression	[28,29]
	Stochastic	Survival function	[14,24,30,31]
		Ordinal regression	[24,32–35]
		Markov chains	[19,28,29,36–40]
		Semi-Markov chains	[24,41,42]
		Discriminant analysis	[19,24]
Data-based	Artificial intelligence	Artificial neural networks — ANNs	[19,24,43–45]
		Fuzzy set	[46–49]
		Case based reasoning — CBR	[50]
		Support vector machines — SVMs	[51]
	Genetic programing	Evolutionary polynomial regression — EPR	[52–54]