Reliability Engineering and System Safety 137 (2015) 120–128

Contents lists available at ScienceDirect

Reliability Engineering and System Safety

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

Computing confidence and prediction intervals of industrial equipment degradation by bootstrapped support vector regression

Isis Didier Lins [a,b,*], Enrique López Droguett [c], Márcio das Chagas Moura [a,b], Enrico Zio [d,e], Carlos Magno Jacinto [f]

[a] Center for Risk Analysis and Environmental Modeling, Federal University of Pernambuco, Recife, Brazil
[b] Department of Production Engineering, Federal University of Pernambuco, Recife, Brazil
[c] Center for Risk and Reliability, Mechanical Engineering Department, University of Maryland, College Park, USA
[d] Department of Energy, Polytechnic of Milan, Milan, Italy
[e] European Foundation for New Energy, Electricité de France, Ecole Centrale Paris – Supelec, France
[f] CENPES, PETROBRAS, Rio de Janeiro, Brazil

ARTICLE INFO

Article history:
Received 4 June 2013
Received in revised form
26 December 2014
Accepted 15 January 2015
Available online 23 January 2015

Keywords:
Degradation
Prognostics
Uncertainty
Confidence and prediction intervals
Support vector regression
Bootstrap
Scale growth rate

ABSTRACT

Data-driven learning methods for predicting the evolution of the degradation processes affecting equipment are becoming increasingly attractive in reliability and prognostics applications. Among these, we consider here Support Vector Regression (SVR), which has provided promising results in various applications. Nevertheless, the predictions provided by SVR are point estimates whereas in order to take better informed decisions, an uncertainty assessment should be also carried out. For this, we apply bootstrap to SVR so as to obtain confidence and prediction intervals, without having to make any assumption about probability distributions and with good performance even when only a small data set is available. The bootstrapped SVR is first verified on Monte Carlo experiments and then is applied to a real case study concerning the prediction of degradation of a component from the offshore oil industry. The results obtained indicate that the bootstrapped SVR is a promising tool for providing reliable point and interval estimates, which can inform maintenance-related decisions on degrading components.

1. Introduction

The prediction of equipment reliability measures is of great importance in many industrial sectors, due to the fact that equipment failures may lead to accidents and/or plant unavailabilities which can translate into costs and production losses [1]. For example, in the context of the oil industry, scale deposition due to salt accumulation may prevent equipment from properly actuating; this may cause the interruption of oil production leading to significant economical losses. Then, the anticipation of potential failures is very attractive because it can enable the implementation of preventive maintenance actions so as to avoid the failures and the associated undesired consequences, costs and losses.

Failure anticipation can be sought by the observation of factors that influence equipment reliability [2,3]. Coming back to the example of scaling build-up in subsea oil well equipment, this process is correlated to a set of explanatory variables, such as reservoir temperature, pressure and water composition, that characterize the subsea environment. These factors can be tracked to predict the amount of scale that will be deposited in the future and determine the time to next maintenance action for removing the scale layer before it builds up to a level that makes the equipment fail to perform its function [2].

Kinetic formulas have been introduced to predict scale formation by correlation of the relevant explanatory variables [4–6]. These deterministic mathematical formulae correlate the scale output variable to the relevant explanatory input variables. Establishing and validating such formula requires a large number of reliable data and is not easy in practice. Lifetime distributions and stochastic processes are also used to model the failure behavior of equipment [7], but generally require making simplifying assumptions rarely met in practice.

Another alternative is to use data-driven learning methods for predicting the evolution of the degradation process affecting the equipment. These methods do not require specific knowledge on the functional relationship between the influential factors (the explanatory variables) and the degradation variable of interest. Among these methods, Support Vector Regression (SVR) has provided promising results in reliability [8,9], economic [10,11],

* Corresponding author.
E-mail address: isis.lins@ceerma.org (I.D. Lins).

http://dx.doi.org/10.1016/j.ress.2015.01.007
0951-8320/

Acronyms

bag, bagging	bootstrap aggregating
CI	confidence interval
ANN	artificial neural network
KKT	Karush–Kuhn–Tucker
LHS	latin hypercube sampling
MSE	mean squared error
PE	point estimate
PI	prediction interval
PSO	particle swarm optimization
RBF	radial basis function
SVR	support vector regression
TBF	time between failures

Notation

B	number of bootstrap samples
b	linear coefficient, bootstrap sample index
C	trade-off between machine capacity and training error
D	training set
$f(\cdot)$	regression function
f_0^0	model trained over original D
$i_1, \ldots, i_\ell$	indexes uniformly generated with replacement from $1, \ldots, \ell$
$K(\cdot)$	kernel function
ℓ	number of training examples
l	observation index
M	number of bootstrap samples of prediction errors
m	index of bootstrap sample of prediction errors
n	number of test examples
T	transpose operator
$\mathbf{w}$	vector of weights
x	input variable
$\mathbf{x}$	input vector
Y	response random variable
y	response observation
$\hat{y}$	estimate of the response variable
α	significance level, Lagrange multiplier
α^*	Lagrange multiplier
δ	prediction error
$\hat{\delta}$	prediction error estimate
ε	tube width
$\epsilon(\cdot)$	random error
$\hat{\epsilon}$	residual
$\tilde{\epsilon}$	re-centered residual
$\check{\epsilon}$	$\tilde{\epsilon}$ modified by η
γ	Gaussian RBF parameter
$\mu_Y(\cdot)$	true mean response
$\boldsymbol{\phi}(\cdot)$	mapping function
$\sigma^2(\cdot)$	variance
η	Rademacher variable
ξ, ξ^*	slack variables
0	subscript indicating optimality
$+$	new observation symbol

environmental [12,13], electrical [14,15] applications, among others. The SVR learning (or training) phase involves the resolution of a convex quadratic optimization problem, for which the Karush–Kuhn–Tucker (KKT) first order conditions are necessary and sufficient for a global optimum [16]. Indeed, this is an advantage of SVR over other learning techniques, such as artificial neural networks (ANNs), which may be trapped into local optima [17]. After the training step, in correspondence of a new observation of the input vector $\mathbf{x}$, hence forth called $\mathbf{x}_+$, the estimate $\hat{y}_+$ of the true mean response $\mu_Y(\mathbf{x}_+)$ can be obtained via the adjusted regression function (*i.e.* the estimator) [17,18].

In many applications, particularly reliability and failure prediction ones, it is important to account for the variability of the estimator and/or assess the uncertainty on the prediction of the response variable Y_+. This means that besides point estimates, confidence intervals for $\mu_Y(\mathbf{x}_+)$ and/or prediction intervals for Y_+ need to be calculated [18,19].

Given that SVR does not require any hypothesis about the distribution of the error term, the central limit theorem enables the approximation of confidence and prediction intervals when large data sets are available [20]. On the other hand, for small numbers of data points (e.g. 100 or less), the intervals based on bootstrap [21,22] tend to be more accurate, given that they do not rely on asymptotic results but on the construction of the limit distribution from the available data [20].

The main idea of bootstrap methods is to estimate probability distributions for statistics of interest obtained from the available data. For regression, bootstrap can be based on: (i) pairs, when the observations from the available data set are sampled with replacement; (ii) on residuals, when the residuals associated with a regression model adjusted over the original sample are sampled with replacement. They are widely used in (generalized) linear, non-linear and nonparametric regression [23]. For example, in linear regression, Cribari-Neto [24] and Cribari–Neto and Lima [25] use bootstrapped hypothesis testing and intervals tailored to account for heteroskedasticity with an estimator of the covariance matrix that considers the effects of leverage points in the design matrix. In nonparametric regression, Zio [26], Cadini et al. [27], Secchi et al. [28] and Zio et al. [29] analyze by bootstrap the uncertainty of ANN predictions of nuclear process parameters. In the context of support vector machines, bootstrap approaches have been mainly applied to classification problems [30,31]. Specifically for SVR, Lin and Weng [32] and Yang and Ong [33] have proposed probabilistic outputs, but assuming a probability distribution for the response variable.

De Brabanter et al. [20] develops approximate pointwise and simultaneous intervals for Least Squares Support Vector Machines (LS-SVM). The authors then compare the approximate intervals with intervals obtained via bootstrap based on residuals (for both homoskedastic and heteroskedastic errors). In the present work, differently from [20], we also consider bootstrap based on pairs, and the SVR procedure put forward involves Vapnik's ε-insensitivity loss function [34] instead of quadratic errors as in LS-SVM.

Lins et al. [35] present a comprehensive approach for variable selection, parameter tuning and uncertainty analysis based on the coupling of Particle Swarm Optimization (PSO), bootstrap methods and SVR. They present an application example for the prediction of Times Between Failures (TBFs) of onshore oil wells located in the Northeast of Brazil. However, in [35], the authors do not validate the combination of bootstrap with SVR for uncertainty characterization and quantification.