Reliability Engineering and System Safety 132 (2014) 9–19

Contents lists available at ScienceDirect

Reliability Engineering and System Safety

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

Mapping model validation metrics to subject matter expert scores for model adequacy assessment

Kirubel Teferra [a,*,1], Michael D. Shields [a,1], Adam Hapij [b], Raymond P. Daddazio [b]

[a] *Department of Civil Engineering, Johns Hopkins University, MD, USA*
[b] *Weidlinger Associates, New York, NY 10005, USA*

ARTICLE INFO

Article history:
Received 6 December 2013
Received in revised form
4 May 2014
Accepted 9 July 2014
Available online 17 July 2014

Keywords:
Model validation
Validation metrics
Subject matter experts
Probabilistic neural networks
Machine learning

ABSTRACT

This paper develops a novel approach to incorporate the contributions of both quantitative validation metrics and qualitative subject matter expert (SME) evaluation criteria in model validation assessment. The relationship between validation metrics (input) and SME scores (output) is formulated as a classification problem, and a probabilistic neural network (PNN) is constructed to execute this mapping. Establishing PNN classifiers for a wide variety of combinations of validation metrics allows for a quantitative comparison of validation metric performance in representing SME judgment. An advantage to this approach is that it semi-automates the model validation process and subsequently is capable of incorporating the contributions of large data sets of disparate response quantities of interest in model validation assessment. The effectiveness of this approach is demonstrated on a complex real-world problem involving the shock qualification testing of a floating shock platform. A data set of experimental and simulated pairs of time history comparisons along with associated SME scores and computed validation metrics is obtained and utilized to construct the PNN classifiers through K-fold cross validation. A wide range of validation metrics for time history comparisons is considered including feature-specific metrics (phase and magnitude error), a frequency metric (shock response spectra), a time-frequency metric (wavelet decomposition), and a global metric (index of agreement). The PNN classifiers constructed using a Parzen kernel for the class conditional probability density function whose smoothing parameter is optimized using a genetic algorithm performs well in representing SME judgment.

1. Introduction

Verification and Validation (V&V) is a formalized methodology to systematically ensure a degree of confidence in a model [14,29–31,38,43,46,51]. Model validation is defined as [29] "*The process of determining the degree to which a model is an accurate representation of the real world from the perspective of the intended uses of the model.*" The model validation process consists of a hierarchy of comparisons between computational models and experiments of system components (subsystems) in order to determine the accuracy of a computational model of the full system. It is often the case that no or minimal experimental data is available for the full system and model error must be quantified through extrapolation. Thus, the model validation process lends itself well to statistical decision-theoretic methods [2], and a framework has been developed based on Bayesian networks [23]. Related concepts of decision theory have been applied extensively for risk-based decision optimization of civil infrastructure systems and networks (e.g. [1,7,8,27,49]). In the Bayesian network framework for the model validation hierarchy [23], the full system is directly represented by a performance function determined by a series of conditional probabilities of dependent nodes. In order to establish the contribution and acceptable errors of the dependent nodes, multiple Monte Carlo simulations of the performance function must be run (i.e. this equates to Monte Carlo simulation of the full system).

Model validation of a subsystem involves evaluating the discrepancy between experimentally recorded and simulated system response quantities (SRQs) of interest- which typically are scalars (random variables) or time histories (random processes)- in order to determine the degree of accuracy of a model. Validation metrics, which are derived from a distance function of a metric space, provide a quantitative goodness-of-fit measure between experimental and simulated SRQs. Definitions and desired properties can be found in [28]. Numerous types of validation metrics have been developed such as probabilistic-based comparisons of scalar quantities (e.g. the area metric [39], Bayesian hypothesis

* Corresponding author. Tel.: +1 212 367 2825.
E-mail address: kteferr1@jhu.edu (K. Teferra).
[1] Author's contributions were made while a research engineer at Weidlinger Associates.

http://dx.doi.org/10.1016/j.ress.2014.07.010
0951-8320/

testing based metrics [16,17,21,23], margin-to-uncertainty based metric [13], reliability-based metric [36]), global comparisons of time histories (e.g. maximum and relative error metrics [28], index of agreement [52], coefficient of efficiency [20]), feature-specific comparisons of time histories (e.g. phase error metrics [10,48], magnitude metrics [10,44]), metrics applied to derived quantities (e.g. wavelet transformations [18], shock response spectra [9]), and many more.

It is emphasized that the acceptance criteria (adequacy) for a subsystem is not only based on validation metric values (measures of accuracy) but also on the contribution of all subsystems to the full system. This is difficult to quantify a priori not only due to computation costs but also due to the lack of understanding of model form error during preliminary full system simulations. Subject matter experts (SMEs) play a vital role in making an interim assessment regarding whether or not a model for a subsystem ought to be accepted or if an alternative model must be studied. Here, the SME makes an adequacy assessment based on the accuracy measured from validation metrics, graphical comparisons of time histories, and expert judgment with regard to the contribution of the subsystem to the performance criteria of the full system. These decisions are associated with large costs as computational models can be expensive for subsystems, especially when uncertainty quantification (UQ) is required.

For complicated physics problems, such as structures subjected to extreme dynamic loading, validation metrics defined over entire time histories are considered necessary (i.e. capturing the salient features of response time histories into a small subset of parameters is an open issue). However, establishing a single or an ensemble of validation metrics for time history comparisons that consistently agree with SME graphical comparisons is very challenging. This is due to both the difficulty in identifying the inherent preferences of SMEs [22], and the imperfect performance of validation metrics for time history comparisons. A study by Schwer [45] compared an ensemble of validation metrics to SME scores and trends in the metrics' performances were observed. A similar study was conducted by Russell [40,41] where the performance of a large number of validation metrics was evaluated for known functions and shock response data. In these studies, comparison of validation metrics to SME opinion is done qualitatively. However, in Ref. [44] linear regression models relating validation metrics to SME scores are fitted and shown to predict SME evaluation (within the variability of SME evaluations) for automobile crash-worthiness tests. The contributions of the current work can be viewed as an extension of Ref. [44]: namely (1) a general framework to incorporate SME evaluation and validation metrics into model validation assessment is proposed and (2) the validation metrics to SME mapping is done through a supervised, nonlinear learning algorithm, which has the capability to improve as more data becomes available.

Establishing a relationship between validation metrics and SME evaluation criteria has the potential to allow for a quantitative performance assessment of validation metrics, identify preference of an SME, and facilitate handling large data sets of validation studies through process semi-automation (e.g. the manual nature of SME evaluation limits the assessment to only spot-checking a small subset of the available time history comparisons.). Tackling these issues is predicated on the existence of a reliable mapping and is beyond the scope of this paper. Instead, this paper focuses on establishing such a mapping by formulating the relationship between validation metrics and SME scores as classification problem, which is then constructed using a probabilistic neural network (PNN) [47]. PNNs are pattern recognition algorithms widely used in machine learning applications representing the human-machine interface [3]. PNNs are constructed for multiple combinations of validation metrics providing both the threshold values of validation metrics corresponding to model accuracy levels (consistent with SME judgment) and a quantitative basis for evaluating their performance. The paper begins with a brief discussion of validation metrics in Section 2 and SME roles in V&V in Section 3. A formulation to determine a model validation score by combining SME scores of various SRQs derived from validation metric values is given in Section 4. A description of PNNs and the particular construct used in this work is in Section 4.1. An example is given in Section 5 where the methodology is exercised on a complex, real-world problem in order to demonstrate its efficacy. The paper concludes with a discussion of the results and summarizes the findings.

2. Discussion of a model validation metric

A metric has been defined as follows [53]: given any three random variables $X_1, X_2, X_3 \in \Omega$ where Ω is a sample space with Borel σ-algebra as the collection of all events $\mathcal{F}$ with measurable probability P, and space $\Omega_2 = \Omega \times \Omega$ being the space of the joint distribution of (X_j, X_k) $\forall j, k$, then metric $M : \{X_j, X_k\} \to \mathbb{R}^+ (\mathbb{R}^+ = [0 \ldots \infty))$ has the following properties:

$$P(X_1 = X_2) = 1 \Rightarrow M(X_1, X_2) = 0, \tag{1a}$$

$$M(X_1, X_2) = 0 \Rightarrow P(X_1 \in f) = P(X_2 \in f), \quad f \in \mathcal{F}, \tag{1b}$$

$$M(X_1, X_2) = M(X_2, X_1), \tag{1c}$$

$$M(X_1, X_3) \le M(X_1, X_2) + M(X_2, X_3). \tag{1d}$$

Thus a metric is a distance function on metric space Ω_2 with the properties of non-negativity (Eq. (1a)), identity (Eq. (1b)), symmetry (Eq. (1c)) and triangular inequality (Eq. (1d)). Metrics applied to the marginal probability density function (PDF) of random variables are denoted as simple metrics, that is, $M : \{f_{X_1}(x_1), f_{X_2}(x_2)\} \to \mathbb{R}^+$, for $f_{X_1}(x_1)$ and $f_{X_2}(x_2)$ being the marginal PDF of random variables X_1 and X_2, respectively. Metrics applied to all other functions of X_1 and X_2 are denoted as compound metrics (this paper is solely concerned with compound metrics; specifically, time histories which can be either realizations or moments of the marginal distributions of random processes).

Consider a computational model $\mathcal{C} : \{p, \theta_{\mathcal{C}}\} \to Y$ that maps input loading $p(\mathbf{x}, t, \theta_{\mathcal{C}})$ to output response $Y(\mathbf{x}, t)$ as a function of spatial and time coordinates $\mathbf{x} \in \mathbb{R}^3, t \in \mathbb{R}^+$, respectively. The uncertainty is modeled by a set of random variables $\theta_{\mathcal{C}}$ representing uncertainty of input loading, model parameters, and model $\mathcal{C}$ itself. This computational model is intended to represent an experimentally tested physical system $\mathcal{P} : \{p, \theta_{\mathcal{P}}\} \to Y$ with p, Y defined as above and $\theta_{\mathcal{P}}$ representing uncertainties associated with the experimental test set up, such as measurement errors and human error.

A model validation metric for time history comparisons of model $\mathcal{C}$, $M : \{Y_k^{(\mathcal{C})}(t), Y_k^{(\mathcal{P})}(t)\} \to \mathbb{R}^+$, is a quantitative measure of the goodness-of-fit between computationally simulated and experimentally recorded SRQs $Y_k^{(\mathcal{C})}(t)$ and $Y_k^{(\mathcal{P})}(t)$, respectively at locations $\mathbf{x}_k : k = 1, 2, \ldots, N_k$ for N_k being the number of locations with sensor recordings. The mathematical definition of a metric (Eqs. (1a)–(1d)) is often loosened in specific application areas so as to obtain a desired performance [11,35]. For example, model validation metrics are typically scaled to be defined in [0 1] and are functions of covariances of X_1,X_2 (violating Eq. (1d)), and are often taken to be relative to the experimental data (violating Eq. (1c)). Specific metrics considered in this work are given in Section 5.1.

3. Model assessment using SME judgment

An SME is broadly defined as someone who has a background in the subject of concern and has qualifications recognized by the