Structural Safety 60 (2016) 77–90

Contents lists available at ScienceDirect

Structural Safety

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

Conditional entropy and value of information metrics for optimal sensing in infrastructure systems

Carl Malings *, Matteo Pozzi

Dept. of Civil and Environmental Engineering, Carnegie Mellon University, Pittsburgh, PA 15213, United States

ARTICLE INFO

Article history:
Received 19 December 2014
Received in revised form 9 October 2015
Accepted 12 October 2015
Available online 26 February 2016

Keywords:
Bayesian analysis
Structural health monitoring
Seismic risk

ABSTRACT

Optimal allocation of monitoring efforts is necessary to cost-effectively obtain information to support the management of civil infrastructure. To optimize the design of sensing networks, pre-posterior analysis of the network can be conducted based on some metric for comparing alternative monitoring schemes. One such metric is conditional entropy, an information theoretic measure of the uncertainty in a set of random variables, conditioned on available sensor measurements. A second metric is the value of information, a decision theoretic metric which explicitly quantifies the benefit of sensor measurements in reducing the expected losses to a managing agent in the context of a decision-making problem under uncertainty. In this paper, we present a scalable probabilistic framework to perform pre-posterior analysis in large infrastructure systems using either metric. A discussion is also provided concerning situations in which either metric should be preferred. To demonstrate this framework, an example infrastructure monitoring problem related to seismic risk is presented and analyzed.

1. Introduction

Management and maintenance of civil infrastructure systems relies on accurate and timely information about the states of the various components which comprise them. Collection of this information can be accomplished by routine inspections, investigations of faults in the systems, or by continuous monitoring of the system through the use of automated sensor networks. Novel sensor networks allow for an unprecedented quantity and quality of information to be available to infrastructure managers, but also may be costly and time-consuming to implement. Therefore, it is worthwhile to investigate the optimal deployment of these sensors, to maximize the potential benefit of the information they would provide while minimizing the sensing costs.

At the individual infrastructure component level, optimal placement of sensors for structural health monitoring (SHM) of bridges or buildings has been investigated for determining the dynamic characteristics of these structures [1–5]. At the infrastructure system level, some investigations have been made into optimal sensing in specific contexts, such as the placement of sensors for contamination detection in water supply networks [6]. In general, sensor placement is a combinatorial optimization problem, in which some subset of objects (sensor locations) is selected from a larger set (all possible locations) in order to optimize some objective (e.g. maximizing sensor coverage or minimizing fault detection time). Such problems can be computationally intensive, and in many cases the only guarantee of finding an optimal solution is to enumerate and compare all possible combinations, which can be intractable in all but the simplest problems. However, heuristics and approximations do exist to overcome this problem [7].

In order to compute metrics for sensor placement, it is necessary to conduct a pre-posterior analysis of proposed sensor networks, such that the potential impact of information gathered by these networks can be predicted before the networks are implemented. The use of pre-posterior analysis for optimal decision-making in civil infrastructure is discussed in [8]. For this analysis to be performed, it is first necessary to have a model defining the possible observations of the sensor network. In this paper, we make use of a Probabilistic Graphical Model (PGM) of the underlying infrastructure system in order to conduct this pre-posterior analysis. PGMs combine random variables, used to quantify the epistemic and aleatory uncertainty in various parameters of the system, with a graphical structure encoding the probabilistic or deterministic relationships between these variables. PGMs are often quite straightforward to interpret visually, and a large variety of PGMs have been designed with the help of experts in various fields to model the behavior of many systems of interest.

* Corresponding author.
E-mail addresses: cmalings@andrew.cmu.edu (C. Malings), mpozzi@cmu.edu (M. Pozzi).

http://dx.doi.org/10.1016/j.strusafe.2015.10.003
0167-4730/

Parameters of these models can be calibrated with the help of expert knowledge, or trained from empirical data. For more information on PGMs, the reader is referred to [9].

PGMs have been used to address issues of both sensor placement and infrastructure management decision-making. In [7], a Gaussian process PGM is used to optimize the placement of sensors for measuring various distributed phenomena, such as temperature and rainfall. In [10,11], a Bayesian network PGM is used to perform risk assessment and decision-making for infrastructure under seismic risk. The basic scheme used to model infrastructure systems under seismic risk in [11] is adapted and used in this paper. The relevant parts of this model will be described in more detail below. A similar model has been employed in [12].

In Section 2, a formal statement of the problem of optimal sensor placement for infrastructure monitoring is presented. In Section 3, the conditional entropy metric is presented. This general metric has been used for guiding optimal sensor placement in many domains where reducing uncertainty is the sensing goal. In Section 4, an alternative metric, the value of information, is also presented. This metric is defined in the context of decision-making problems, and can be used to guide optimal sensor placements for supporting this decision-making. We propose a general method for computing and optimizing these metrics in Section 5. In Section 6, a scheme for representing infrastructure systems and conducting inference using multivariate Gaussian models is developed. In Section 7, we describe the application of these metrics and models to infrastructure systems subjected to seismic risks. The use of these metrics for optimizing sensor placement is demonstrated in Section 8 for an example application involving a network of bridges and tunnels in the San Francisco Bay area. Further investigations into the densities of these metrics over the seismic risk scenarios are presented in Section 9, before drawing conclusions in Section 10.

2. Problem statement

Consider an infrastructure system consisting of n components. These components are subjected to a risk scenario $s \in S$, referring to a specific type of loading placed upon the system, e.g. an earthquake, where S is the domain of possible risk scenarios. A set of variables $W = \{W_1, \ldots, W_n\}$ describes all components within the system, where variable W_i, describing component i, can be multidimensional, so as to incorporate all relevant features which characterize the component. In the context of spatially distributed infrastructure systems, W refers to any set of structural component properties, as well as external factors such as the loads or demands, to which these components are subjected. In this paper, we restrict ourselves to considering scalar quantities, e.g. the maximum loads on a structure due to an extreme event (as is considered in the example presented in Sections 7 and 8).

Some deterministic function of features W defines the states of the components, in terms of their behavior or functionality. We assume that the state of each component is binary, i.e., the component can either be in an "operational" or "failed" state. We describe with binary variable X_i the state of component i, and denote by X the set of $X_1, \ldots, X_n$, the joint states of all components in the system.

Through inspections or sensor measurements, we may observe some subset of the features W. We denote by Ω the set of all such possible observations, with $\Omega = \{\Omega_1, \ldots, \Omega_n\}$, where Ω_i denotes all possible observations of features related to component i. These observations are potentially noisy or indirect measures of the features in W. Note that we assume that variables within X cannot be observed directly (unless perfect measurements are made of all features of the component relevant to determining its state); this assumption is made to avoid computational difficulties within the framework presented here, as discussed in Section 6. Sensor placement involves the selection of some set Y of variables to observe, where $Y = \{Y_1, \ldots, Y_n\}$ with $Y_i \subseteq \Omega_i$. Note that Y_i may be an empty set if no features relating to component i are selected for observation by sensors. This formulation can also easily be generalized to include the possibility of collecting measurements where no components are located by introducing "pseudo-components" endowed with observable features but whose states are already known with certainty, e.g. "operational".

Fig. 1 shows a general PGM for a system as outlined above. The representation follows a common convention, with circles indicating random variables and arrows or lines indicating the dependence structure among these variables. Shaded circles indicate observed variables [9]. As reported in the figure, we assume that, although a priori the scenario S is uncertain, when a scenario occurs, it can be directly identified. Therefore, we assume that sensor placement is performed under uncertainty on S, however, processing of data collected by sensors in the aftermath of the scenario will be conditioned upon knowledge of S. This is a reasonable assumption in certain domains, such as seismic risk, where earthquake scenario information (the epicenter location and magnitude) are usually determined with a high degree of accuracy in the aftermath of the scenario.

Within this PGM, the problem of optimal sensor placement can be formalized as follows:

$$Y^* = \operatorname{argmax}_{Y \subseteq \Omega} m_X(Y) \quad \text{subject to} \quad \mathrm{C}(Y) \leqslant B \tag{1}$$

where Y^* is the optimal set of sensed variables, $\mathrm{C}(Y)$ is the cost of implementing a sensing network to measure Y, B is a fixed budget constraint, and $m_X(Y)$ is a metric which quantifies how observing Y will improve knowledge of the state X of the components in the system. Note that, while W gives a full description of the system, we seek to place sensors to optimally gain knowledge about X, rather than W, since performance of the components, captured in X, is the true information of interest in managing the system.

We model the uncertain scenario $s \in S$ through distribution $\mathrm{p}(S)$. If $m_{X|s}(Y)$ is our sensor placement metric, computed under a specified scenario s, then we can compute an expected value of $m_{X|s}(Y)$ over $\mathrm{p}(S)$ to obtain the metric value $m_X(Y)$ across all risk scenarios:

$$m_X(Y) = \mathbb{E}_S\left[m_{X|s}(Y)\right] = \int_S m_{X|s}(Y)\,\mathrm{p}(S)\,\mathrm{d}s \tag{2}$$

where $\mathbb{E}_S[\cdot]$ indicates the statistical expectation under the probability distribution $\mathrm{p}(S)$ of S.

In this paper, we discuss two ways to define the metric $m_{X|s}(Y)$: conditional entropy and the value of information. The conditional entropy metric, as presented in Section 3, addresses the problem

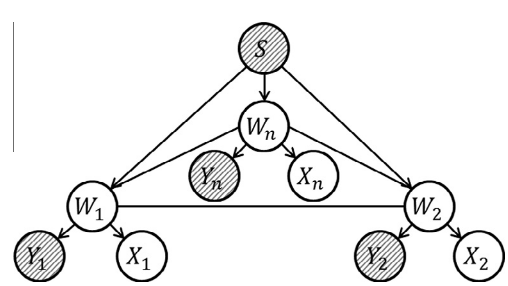

Fig. 1. A general probabilistic graphical model for a partially observable system with n components. Variable S describes the risk scenario. Variable W_i describes component i, binary variable X_i gives its state, and variable Y_i is a selected measurement of observable features for this component.