Engineering Applications of Artificial Intelligence 70 (2018) 1–15

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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

A local multiscale probabilistic graphical model for data validation and reconstruction, and its application in industry

Javier Herrera-Vega [a,*], Felipe Orihuela-Espina [a], Pablo H. Ibargüengoytia [b], Uriel A. García [b], Dan-El Vila Rosado [a], Eduardo F. Morales [a], Luis Enrique Sucar [a]

[a] Instituto Nacional de Astrofísica Óptica y Electrónica, Puebla, Mexico
[b] Instituto de Investigaciones Eléctricas, Cuernavaca, Mexico

ARTICLE INFO

Keywords:
Bayesian networks
Data validation
Multiscale approach
Outlier detection
Probabilistic graphical models

ABSTRACT

The detection and subsequent reconstruction of incongruent data in time series by means of observation of statistically related information is a recurrent issue in data validation. Unlike outliers, incongruent observations are not necessarily confined to the extremes of the data distribution. Instead, these rogue observations are unlikely values in the light of statistically related information. This paper proposes a multiresolution Bayesian network model for the detection of rogue values and posterior reconstruction of the erroneous sample for non-stationary time-series. Our method builds local Bayesian Network models that best fit to segments of data in order to achieve a finer discretization and hence improve data reconstruction. Our local multiscale approach is compared against its single-scale global predecessor (assumed as our gold standard) in the predictive power and of this, both error detection capabilities and error reconstruction capabilities are assessed. This parameterization and verification of the model are evaluated over three synthetic data source topologies. The virtues of the algorithm are then further tested in real data from the steel industry where the aforementioned problem characteristics are met but for which the ground truth is unknown. The proposed local multiscale approach was found to dealt better with increasing complexities in data topologies.

1. Introduction

Many areas like industry, medicine and science generate large volumes of data demanding validation. Validating data is a crucial task before information analysis, interpretation and decision making. Data validation encompasses processing techniques rendering quality data guaranteeing optimal matching between real observations and the repository. In other words, data validation is concerned with finding erroneous data in a data set and when appropriate, suggesting a plausible alternative (Tamrapani and Johnson, 2003). The data validation process involves a systematic assessment of compliance to a set of acceptance rules defining data validity (Herrera-Vega et al., 2012). In general, the validation process is domain specific (Gonzalez et al., 2012; Lamrini et al., 2011). Due to its domain specific nature, data validation is carried out not in few occasions by means of visual inspection, a time consuming approach, exposed to subjectiveness and prone to errors. Yet regardless of the particularities in each domain, a number of problems are recurrent in data validation including *detection* of outliers, incongruent or rogue values and/or gaps or missing data, and *reconstruction* or estimation of these missing or erroneous observations (Ibargüengoytia et al., 2013).

For some of these common data validation problems automation has been attempted (Abraham and Box, 1979; Blake, 1993; Bao and Dai, 2009; Hoo et al., 2002; Peng et al., 2012; Tsay, 1988; Walczak, 1995). These problems include the detection of observational outliers (values at the extreme of the data distribution), the detection of signal drift, level shift or abrupt changes altering the trend of the series (innovation outliers Muirhead, 1986; Abraham and Box, 1979), the detection of rogue values (unplausible values in the light of statistically dependent information) and the reconstruction of missing data, among others. Once an error has been detected, the validation process proceeds with the data reconstruction. Reconstruction can capitalize on the signal autoregressive information e.g., classical interpolation (Stoer and Bulirsch, 2002) or time series analysis (Box et al., 2013), statistically dependent information e.g., Lamrini et al. (2011) and Ibargüengoytia et al. (2006), or a combination of both (Ibargüengoytia et al., 2013). The

* Corresponding author.
E-mail addresses: vega@ccc.inaoep.mx (J. Herrera-Vega), f.orihuela-espina@ccc.inaoep.mx (F. Orihuela-Espina), pibar@iie.org.mx (P.H. Ibargüengoytia), uriel.garcia@iie.org.mx (U.A. García), dnvr301080@ccc.inaoep.mx (D.-E. Vila Rosado), emorales@inaoep.mx (E.F. Morales), esucar@ccc.inaoep.mx (L.E. Sucar).

https://doi.org/10.1016/j.engappai.2018.01.001
Received 30 November 2016; Received in revised form 21 November 2017; Accepted 1 January 2018
0952-1976/

most beneficial reconstruction option depends on the interplay between the variables characteristics in the dataset, including the within-variable information (Ibargüengoytia et al., 2013).

This paper is concerned with the detection of incongruent values and its reconstruction paying particular attention to datasets with temporal variables (time series). Incongruent or rogue values are suspicious values which may be in range and apparently agree with the signal trend, but that contradicts the associated trend of statistically dependent knowledge (Herrera-Vega et al., 2012). This makes their detection particularly difficult if using only within-variable information. This is inherently a multivariate problem. Previously, the detection of rogue values has been addressed with Bayesian Networks (BNs) in the context of sensor validation (Ibargüengoytia et al., 2006). This approach is a *global* solution in which a BN is learned from the complete available dataset and then data validity is checked against probabilistic plausibility. Subsequent reconstruction utilizes probabilistic propagation to estimate expected values for erroneous samples from associated values in statistically related variables (Ibargüengoytia et al., 2006). Learning the structure of the BN requires discretization of variables' data ranges into intervals that ultimately determines the detection rate and affects the accuracy of the recovery of alternative values. For stationary signals it is fair that these discrete intervals remain constant for the whole time series. However, for non-stationary signals, as the statistical properties of the series fluctuate, so should the intervals. In this way, dynamic finer discretization can be achieved and consequently a more accurate suggestion of alternative values should follow. Achieving similar discretization with a global solution will imply higher number of intervals, which in turn will require conditional probability tables that will grow exponentially. This quickly becomes computationally intractable.

This paper proposes a new *local* multiscale BN-based approach for the detection and reconstruction of incongruent values in multivariate datasets that we hypothesize to be more suitable for non-stationary signals. The algorithm constructs a two level hierarchy of BN models in which the superior level determines the dataset topology i.e., BN structure, and the inferior level contains a set of submodels providing interval discretizations that locally fits data distribution. In detecting the error and reconstructing the new value, the critical step is deciding the submodel that better fits the sample under scrutiny. The problem of selecting the submodel is solved computing the conditional probability of the observation given the submodel. The solution aims to enhance error detection and suggestion of alternative values that offer a greater congruence with the data series trend by computing conditional probabilities locally. Validation is carried out over synthetic data. Explicative power is evaluated by matching the reconstructed Bayesian structure against the known synthetic ground truth. Predictive power is assessed in its two flavours, error detection capabilities and error reconstruction capabilities and compared against the global predecessor.

The detection process is then applied to a subset of the data coming from the hardening furnace. Manufacturing of seamless steel tubes used for operating at high temperatures and pressures requires the creep resistance resulting from heat treatment. Heat treatment is a set of metalworking processes that alter the mechanical characteristics of the material by means of a sequence of heating and/or cooling to extreme temperatures. For instance, one such heat treatment, annealing, changes material properties such as strength and hardness. This manufacturing process often yields a wealth of data with over 120 different variables with very long series. This data is used to classify the steel tube as compliant or not with resistance requirements. This classification is strongly affected by the quality of the data. However, during data acquisition and storage; defective sensing, noise affecting transmission and transcription mistakes may corrupt the data. Thus, to achieve a more accurate classification, data is put through a data validation process. In this scenario, type I errors i.e., considering faulty an actual correct value, are affordable as long as the suggested alternatives are good approximations, emphasizing the critical importance of the reconstruction. Performance is then compared to the previous existing global solution (Ibargüengoytia et al., 2006).

Contribution is three-fold; (i) we provide a new solution with overall better capabilities for complex data topologies, (ii) we establish some rules of thumb for model parameterization both for the new approach and its predecessor, and (iii) we verify and validate the approach delimiting its incongruent data validation capabilities.

Organization of the paper is as follows. First, the computational approach is presented, and the datasets both synthetic and real are introduced. Then, to reduce the search space, an initial stage chooses statistically relevant model parameters. After fixing the model parameters, a 5-fold validation exercises explores the face validity of the approach evaluating predictive and explicative properties of the solution. The model parameters are automatically learned from the data structure and distribution to adapt to different problems and scenarios, and in principle it can accommodate any number of variables (other than memory limits) and it is not constraint to a particular data distribution favouring scalability and generalizability. Finally, concurrent validity is established over real data.

2. Preliminaries

2.1. Bayesian networks

A Bayesian Network (BN) (Koller and Friedman, 2009; Pearl, 1982) $N = (X, G, P)$ is a directed acyclic graph (DAG) $G = (V, E)$ with nodes $V = v_1, \ldots, v_n$ and directed links E. The nodes of G represent the set of random variables X of the domain and for each random variable $X_v \in X$ a *Conditional Probability Distribution* P of the form $P(X_v|X_{pa}(v))$ is associated, where $X_{pa}(v)$ are the set of parents of v.

The BN structure and its parameters (CPD) can be defined explicitly. However, these can be learned automatically through a set of data. Several algorithms are available for this purpose (Spirtes et al., 2000; Steck, 2001; Chow and Liu, 1968). During the network structure learning, a statistical test between each pair of variables must be computed in order to discover relations of conditional independence, a process that is facilitated by the discretization of the variables' ranges.

A defined (learned) BN represents a knowledge base which its mayor purpose is to reason under uncertainty about observed events in its domain. This reasoning is carried out by probabilistic inference whose main task is to compute the posterior marginal probability of an unobserved variable given a set of observed (evidence) variables.

2.2. Discretization

Discretization is the process by which the values of continuous variables are converted to discretized, ordinal or nominal values. The discretization process is non-trivial and many approaches exist (Kotsiantis and Kanellopoulos, 2006; Friedman and Goldszmit, 1996). Two classical strategies are equi-distance by which the variables' data range is split in a predetermined number of equally distant intervals, or equi-frequency in which the splitting of the intervals ensure that each interval holds the same number of samples. Recently, we proposed an interval discretization technique based on a Gaussian mixture model (GMM) (Herrera-Vega et al., 2012). This approach optimizes binning based on the data distribution. In GMM-based interval discretization, the data is assumed to be generated by a mixture of Gaussian distributions. Each fundamental Gaussian is characterized by its mean μ and its variance σ^2, and the mixture is given by Eq. (1):

$$p(x) = \sum_{k=1}^{K} \pi_k N(x|\mu_k, \sigma_k^2) \tag{1}$$

where K is the number of Gaussians considered, $N(x|\mu_k, \sigma_k^2)$ represents a Gaussian with mean μ_k and variance σ_k^2 and π_k are the mixing coefficients, i.e. weights for the Gaussians. The algorithm has K as a single parameter. The classical Expectation–Maximization algorithm (Dempster et al., 1977) is used to optimize the fitting of the distributions. The critical value discriminating any two contiguous intervals is chosen