Ecosystem Services xxx (2018) xxx–xxx

Contents lists available at ScienceDirect

Ecosystem Services

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

Machine learning for ecosystem services

Simon Willcock [a,b,*], Javier Martínez-López [c], Danny A.P. Hooftman [d,e], Kenneth J. Bagstad [f], Stefano Balbi [c], Alessia Marzo [g], Carlo Prato [g], Saverio Sciandrello [g], Giovanni Signorello [g], Brian Voigt [h], Ferdinando Villa [c,i], James M. Bullock [e], Ioannis N. Athanasiadis [j]

[a] School of Environment, Natural Resources and Geography, Bangor University, United Kingdom
[b] Biological Sciences, University of Southampton, United Kingdom
[c] Basque Centre of Climate Change, Spain
[d] Lactuca: Environmental Data Analyses and Modelling, The Netherlands
[e] NERC Centre for Ecology and Hydrology, Wallingford, United Kingdom
[f] Geosciences & Environmental Change Science Center, U.S. Geological Survey, Denver, CO, USA
[g] Centre for the Conservation and Management of Nature and Agroecosystems (CUTGANA), University of Catania, Catania, Italy
[h] Gund Institute, Rubenstein School of Environment and Natural Resources, University of Vermont, Burlington, VT, USA
[i] IKERBASQUE, Basque Foundation for Science, Bilbao, Bizkaia, Spain
[j] Information Technology Group, Wageningen University, The Netherlands

ARTICLE INFO

Article history:
Received 29 September 2017
Received in revised form 1 March 2018
Accepted 11 April 2018
Available online xxxx

Keywords:
ARIES
Artificial intelligence
Big data
Data driven modelling
Data science
Machine learning
Mapping
Modelling
Uncertainty, Weka

ABSTRACT

Recent developments in machine learning have expanded data-driven modelling (DDM) capabilities, allowing artificial intelligence to infer the behaviour of a system by computing and exploiting correlations between observed variables within it. Machine learning algorithms may enable the use of increasingly available 'big data' and assist applying ecosystem service models across scales, analysing and predicting the flows of these services to disaggregated beneficiaries. We use the Weka and ARIES software to produce two examples of DDM: firewood use in South Africa and biodiversity value in Sicily, respectively. Our South African example demonstrates that DDM (64–91% accuracy) can identify the areas where firewood use is within the top quartile with comparable accuracy as conventional modelling techniques (54–77% accuracy). The Sicilian example highlights how DDM can be made more accessible to decision makers, who show both capacity and willingness to engage with uncertainty information. Uncertainty estimates, produced as part of the DDM process, allow decision makers to determine what level of uncertainty is acceptable to them and to use their own expertise for potentially contentious decisions. We conclude that DDM has a clear role to play when modelling ecosystem services, helping produce interdisciplinary models and holistic solutions to complex socio-ecological issues.

1. Introduction

Many scientific disciplines are taking an increasingly integrative approach to planetary problems such as global climate change, food security and human migration (Baziliana et al., 2011; Bullock et al., 2017). To address such challenges, methods and practices are becoming more reliant on large, interdisciplinary data repositories often collected in cutting-edge ways, for example via citizen scientists or automated data collection (Isaac et al., 2014). Recent developments in information technology have expanded modelling capabilities, allowing researchers to maximise the utility of such 'big data' (Lokers et al., 2016). Here, we focus on one of these developments: data-driven modelling (DDM). DDM is a type of empirical modelling by which the data about a system are used to create models, which use observed systems' states as inputs for estimating some other system state(s), i.e., outputs (Jordan and Mitchell, 2015; Witten et al., 2016). Thus, DDM is the process of identifying useful patterns in data, a process sometimes previously referred to as knowledge discovery in databases (Fayyad et al., 1996). This process consists of five key steps: 1) understanding the research goal, 2) selecting appropriate data, 3) data cleaning, pre-processing and transformation, 4) data mining

* Corresponding author at: School of Environment, Natural Resources and Geography, Bangor University, United Kingdom.

E-mail addresses: s.willcock@bangor.ac.uk (S. Willcock), javier.martinez@bc3research.org (J. Martínez-López), danny.hooftman@lactuca.nl (D.A.P. Hooftman), kjbagstad@usgs.gov (K.J. Bagstad), stefano.balbi@bc3research.org (S. Balbi), alessia.marzo@unict.it (A. Marzo), s.sciandrello@unict.it (S. Sciandrello), g.signorello@unict.it (G. Signorello), Brian.Voigt@uvm.edu (B. Voigt), ferdinando.villa@bc3research.org (F. Villa), jmbul@ceh.ac.uk (J.M. Bullock), ioannis@athanasiadis.info (I.N. Athanasiadis).

https://doi.org/10.1016/j.ecoser.2018.04.004
2212-0416/© 2018 The Authors. Published by Elsevier B.V.

(creating a data driven model), and 5) interpretation/evaluation (Fayyad et al., 1996) (Fig. 1). A variety of methods for data mining and analysis are available, some of which utilise machine learning algorithms (Witten et al., 2016; Wu et al., 2014) (Fig. 1). A machine learning algorithm is a process that is used to fit a model to a dataset, through training or learning. The learned model is subsequently used against an independent dataset, in order to determine how well the learned model can generalise against the unseen data, a process called testing (Ghahramani, 2015; Witten et al., 2016). This training–testing process is analogous to the calibration–validation process associated with many process-based models.

In general, machine learning algorithms can be divided into two main groups (supervised- and unsupervised-learning; Fig. 1), separated by the use of explicit feedback in the learning process (Blum and Langley, 1997; Russell and Norvig, 2003; Tarca et al., 2007). Supervised-learning algorithms use predefined input-output pairs and learn how to derive outputs from inputs. The user specifies which variables (i.e., outputs) are considered dependent on others (i.e., inputs), which sometimes indicates causality (Hastie et al., 2009). The machine learning toolbox includes several linear and non-linear supervised learners, predicting either numeric outputs (regressors) or nominal outputs (classifiers) (Table 1). An example of supervised machine learning that is familiar to many ecosystem service (ES) scientists is using a general linear model, whereby the user provides a selection of input variables hypothesised to predict values of an output variable and the general linear model learns to reproduce this relationship. The learning process needs to be fine-tuned through a process, as for example in the case of stepwise selection where an algorithm selects the most parsimonious best-fit model (Yamashita et al., 2007). However, note that stepwise functions may also be used in unsupervised learning processes when combined with other methods. Within unsupervised-learning processes, there is no specific feedback supplied for input data and the machine learning algorithm learns to detect patterns from the inputs. In this respect, there are no predefined outputs, only inputs for which the machine learning algorithm determines relationships between them (Mjolsness and DeCoste, 2001). An example unsupervised-learning algorithm, cluster analysis, groups variables based on their closeness to one another, defining the number and composition of groups within the dataset (Mouchet et al., 2014). Within the supervised- and unsupervised-learning categories, there are several different varieties of machine learning algorithms, including: neural networks, decision trees, decision rules and Bayesian networks. Others have described the varieties of machine learning algorithms (Blum and Langley, 1997;

Table 1
A simplified summary of machine learning algorithms (categorised as supervised and unsupervised).

Category	Task	Common algorithms
Unsupervised learning (learning without feedback from a trainer)	Clustering	k-means
	Associations	Apriori
	Dimensionality reduction	PCA
Supervised learning (learning past actions/ decisions with trainer)	Classification (learning a categorical variable)	Bayesian Networks, Decision Trees, Neural Networks
	Regression (learning a continuous variable)	Linear Regression, Perceptron

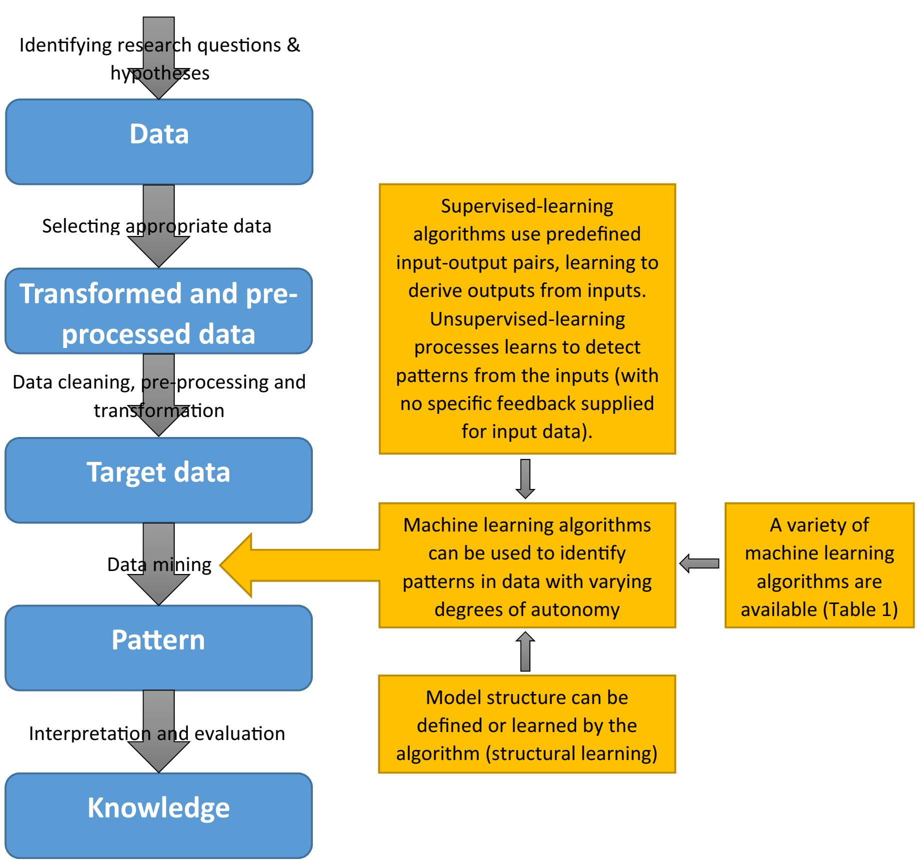

Fig. 1. A schematic outlining how machine learning algorithms (yellow) can contribute to the data-driven modelling process (blue) (Fayyad et al., 1996). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Please cite this article in press as: Willcock, S., et al. Machine learning for ecosystem services. Ecosystem Services (2018), https://doi.org/10.1016/j.ecoser.2018.04.004