Ecological Modelling 368 (2018) 391–403

Contents lists available at ScienceDirect

Ecological Modelling

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

Discretizing environmental data for learning Bayesian-network classifiers

R.F. Ropero[a,*], S. Renooij[b], L.C. van der Gaag[b]

[a] Informatics and Environment Laboratory, Dept. of Biology and Geology, University of Almería, Carretera de Sacramento s/n, La Cañada de San Urbano, Almería, Spain
[b] Department of Information and Computing Sciences, Utrecht University, Princetonplein 5, De Uithof, Utrecht, The Netherlands

ARTICLE INFO

Article history:
Received 18 January 2017
Received in revised form 5 December 2017
Accepted 6 December 2017
Available online 2 January 2018

Keywords:
Species distribution models
Bayesian-network classifiers
Logistic-regression models
Discretization methods

ABSTRACT

For predicting the presence of different bird species in Andalusia from land-use data, we compare the performances of Bayesian-network classifiers and logistic-regression models. In our study, both well balanced and less balanced data sets are used, and models are learned from both the original continuous data and from the data after discretization. For the latter purpose, four different discretization methods, called *Equal Frequency*, *Equal Width*, *Chi-Merge* and *MDLP*, are compared. The experimental results from our species data sets suggest that the simple Naive Bayesian classifiers are preferable to logistic-regression models and that the relatively unknown *Chi-Merge* method is the preferred method for discretizing these environmental data.

1. Introduction

Bayesian networks (BNs for short) are powerful probabilistic models that have demonstrated their usefulness in a wide range of application fields among which is the environmental-science field (Baur and Bozdag, 2015; Jensen and Nielsen, 2007). In environmental science, Bayesian networks are used for knowledge discovery, where the focus is on establishing the relationships among the variables at hand and their evolution under various scenarios (Dyer et al., 2014). Bayesian networks are further used for classification purposes (Maldonado et al., 2015; Park and Stenstrom, 2008), where the aim is to accurately predict the value of a specific target variable, called the class variable.

Initially, Bayesian networks were designed to handle data pertaining to discrete variables only. Real-world data are often of a continuous or hybrid nature however, and new algorithms for learning and inference in Bayesian networks with both continuous and discrete variables are emerging (Langseth et al., 2012; Moral et al., 2001). Despite the increasing availability of such algorithms, most Bayesian-network packages to date require variables to be discrete. Upon practical application, therefore, any continuous variables need to be discretized.

Discretization is widely applied in knowledge-discovery and machine-learning applications, with the aim of *(i)* reducing and simplifying the available data, *(ii)* rendering model learning more efficient, and *(iii)* obtaining more compact and readily interpretable results (Liu et al., 2002). Over the years, several different discretization methods have been proposed, only a few of which are widely used while others are largely unnoticed (García et al., 2013; Yang et al., 2010; Liu et al., 2002). Since data discretization generally results in information loss (Li, 2007; Uusitalo, 2007), the discretization method employed will affect the predictive quality of any model learned from the data. Where several papers address the question of which discretization method is most suited for data mining in general (García et al., 2013; Liu et al., 2002) or for Bayesian-network learning in particular (Lima et al., 2014; Zhou et al., 2014), the best choice of method tends to depend on the nature and characteristics of the data at hand.

In environmental science, Bayesian networks are typically used in a decision-making process in which expert knowledge plays an important role (Voinov and Bousquet, 2010). In this context, the use of discrete data provides more easily interpretable results and facilitates the communication between modelers and environmental experts (García et al., 2013; Liu et al., 2002). According to a recent review (Aguilera et al., 2011), in fact, more than 80% of the papers addressing Bayesian networks in environmental science involve discretized data, where the discretization is done using the so-called *Equal Frequency* method or is based on expert knowledge. While more tailored discretization methods have been

* Corresponding author.
E-mail addresses: rosa.ropero@ual.es (R.F. Ropero), S.Renooij@uu.nl (S. Renooij), L.C.vanderGaag@uu.nl (L.C. van der Gaag).

https://doi.org/10.1016/j.ecolmodel.2017.12.015
0304-3800/

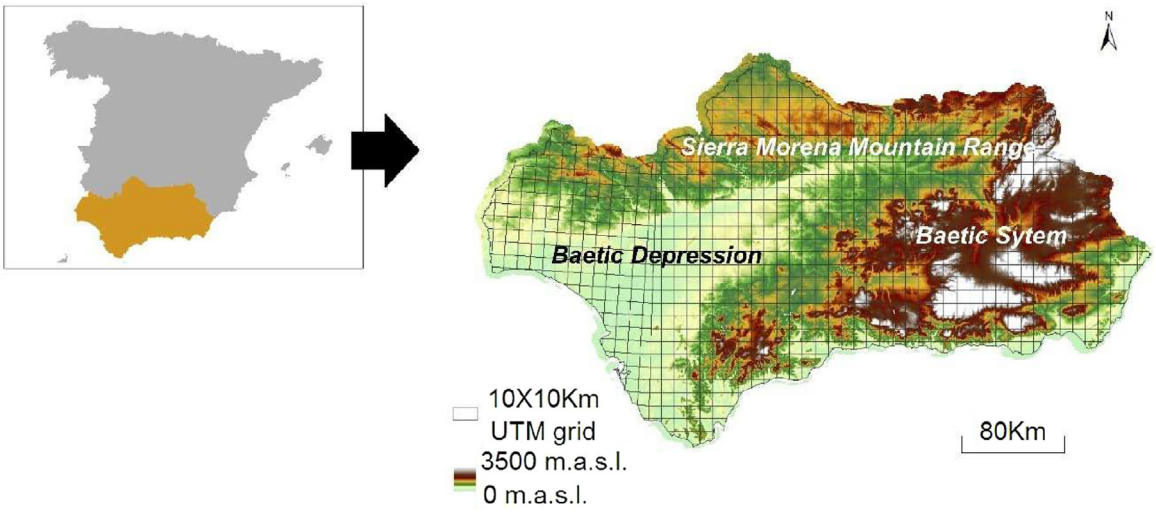

Fig. 1. Andalusia, located in the South of Spain (*left*), its relief and the UTM 10 × 10 km grid used for the data collection (*right*); the smaller cells in the western area result from the grid having been corrected to fit two geographical HUSOS.

designed for specific types of model, such as hydrological models (Pradhanang and Briggs, 2014), models of air quality (Davison and Ramesh, 1996), and models of spatial distributions of the data (Liu et al., 2015), discretization methods specifically designed for environmental modeling through Bayesian networks do not abound. To bring the discretization methods in use with Bayesian networks in general to the attention of environmental modelers, further efforts as well as more tailored insights are called for (Nash et al., 2013).

During the last decades, species distribution modeling has evolved in the field of environmental science, following the development of Geographic Information Systems (GIS) and spatial statistics techniques (Segurado and Araújo, 2004). In general, the objective of species distribution modeling is to link species data with environmental variables and to obtain maps showing the spatial distribution of the species under study (Elith et al., 2006). Some of the most commonly used models for this purpose are classification trees (Fukuda et al., 2013), regression models (Li and Wang, 2013), neural networks (Dedecker et al., 2004), and more tailored models like BIOCLIM (Busby, 1986) and FLORAMAP (Jones and Gladkov, 1999). In contrast, Bayesian networks are scarcely being applied in species distribution modeling, although some examples are found, addressing classification with discretized data (Newton et al., 2007) and using a model structure based on expert knowledge (Pollino et al., 2007).

In this paper we compare various classification models for predicting the presence of different bird species in Andalusia from land-use data. More specifically, we study the performance of two types of Bayesian-network classifier: the Naive Bayesian (NB) classifier and the Tree Augmented Naive Bayesian (TAN) classifier. These classifiers are learned from both the original continuous data and from discretized data. For discretization, four methods are compared: *Equal Frequency* (EF), *Equal Width* (EW), *Chi-Merge* (ChiM) and a method based on the *Minimum Description Length Principle* (MDLP); these methods are the most commonly used discretization methods (García et al., 2013; Liu et al., 2002). We further compare the performances of these classifiers when learned from well balanced data sets and from less balanced data.

The performance of a classification model depends to a large extent on the decision rule that is used to decide upon the class to which a case is assigned. In practice often maximum-probability classification is used, in which a case is assigned to the most likely class (Ropero et al., 2015; Aguilera et al., 2013). In essence, however, any probability can be chosen for a decision threshold: a species then is classified as *present* if the predicted probability of it being present exceeds this threshold, and as *absent* otherwise. For less balanced data sets, in which the prior distribution over the class variable is quite skewed, maximum-probability classification may lead to undesirable classification behaviour (van der Gaag et al., 2009a,b). In this paper we therefore study the performance of the various classifiers with maximum-probability classification and with threshold-probability classification using a decision threshold based on the prior species distribution (van der Gaag et al., 2009a,b).

Since in species distribution modeling the use of logistic-regression models is quite common, from the various data sets also logistic-regression models are constructed and compared with the learned Bayesian-network classifiers in terms of their performance.

2. Materials and methods

In this section we review the data sets used in our study and introduce the various methods for discretizing these data and for learning and validating classification models.

2.1. Study area and data collection

Andalusia, located in the South of Spain (Fig. 1), constitutes the nation's second largest autonomous region, with a surface area of 87,600 km^2 representing 17.3% of the national territory.[1] Lying on the frontier between Europe and Africa, Andalusia inherits landscape and biodiversity specifics from both continents. Its terrain covers a wide range of altitudes, from the Baetic Depression to the mountainous ranges of the Sierra Morena and the Baetic System, with the highest peaks lying over 3000 meters above see level (m.a.s.l.) The landscape is quite heterogeneous, with huge differences from the densely populated and irrigated cropland areas of the river basin and coastlands, to the sparsely populated forested areas of the uplands. Its climate is similarly heterogeneous, with stark differences between inland and coastal areas. The climate in the south-eastern coastal part is semiarid, with less than 200 mm of annual rainfall in several areas, while the middle and northern parts have a continental climate, with more than 4000 mm of rainfall per year. These natural conditions make Andalusia a heterogeneous region both in terms of territorial structure and in climatic and

[1] Data from the Spanish Statistical Institute.