[redacted]

Discretizing environmental data for learning Bayesian-network classifiers

[redacted]

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

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.

[redacted]

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 designed for specific types of model, such as hydrological models

[*] 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).

[redacted]

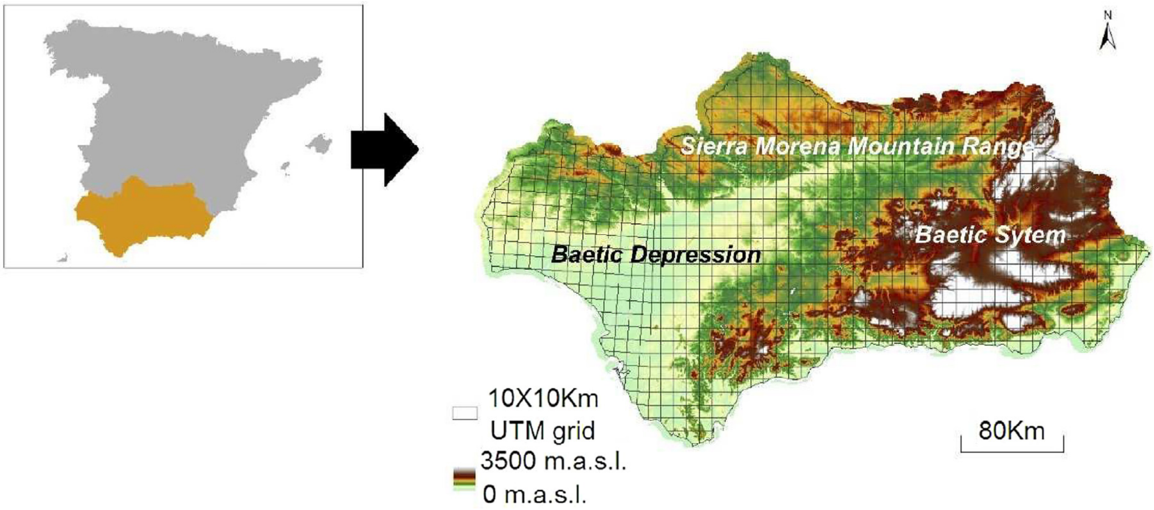


Fig. 1. Andalusia, located in the South of Spain (*left*), its relief and the UTM 10 × 10 km grid used for the data collection (*right*); the smaller cells in the western area result from the grid having been corrected to fit two geographical HUSOS.

(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.

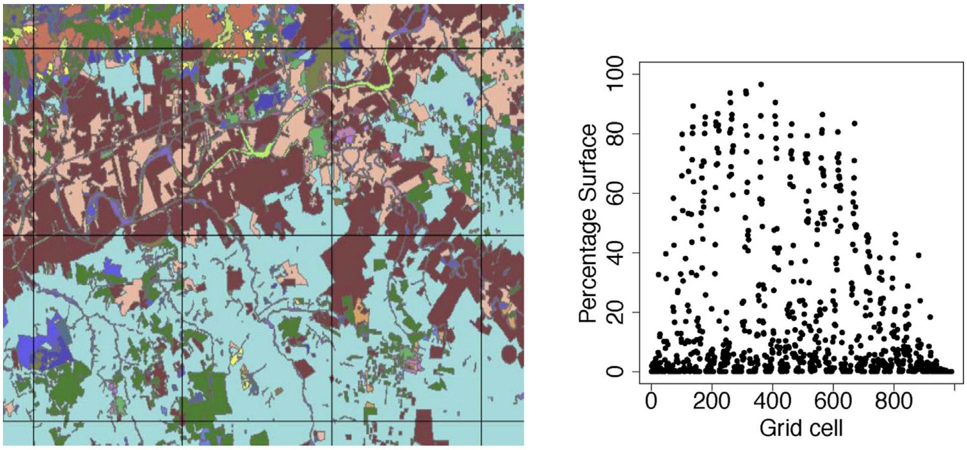
2. Materials and methods

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 ecological conditions. Provoking ecological niches with large biodi-

[1] Data from the Spanish Statistical Institute.



(a)

(b)

Fig. 2. Enlarged part of the 10 × 10 km grid showing different types of land use (a), and the distribution of *Olive cropland* coverage over all grid cells.

versity rates, Andalusia is considered a global biodiversity hotspot (Myers et al., 2000).

The Spanish Inventory of Terrestrial Species[2] by the Spanish National Government was used to select information about the prevalence of three bird species – *Turdus viscivorus*, *Cecropis daurica* and *Accipiter nisus* – for the UTM (*Universal Transverse Mercator*) 10 × 10 km grid of Andalusia (Fig. 1); the three species were selected for their different prevalence rates. Information about land use for the same grid was collected from the Andalusian Environmental Information Network[3] from the Andalusian Regional Government. ArcGIS 9.3 was used for selecting the data and merging them into the grid. As a consequence of the high heterogeneity of the region, a single cell of the grid of Andalusia can show several small patches of different types of land use, as illustrated in Fig. 2(a). A more detailed example, showing the distribution of *Olive cropland* coverage over the grid cells, is provided in Fig. 2(b). The figure shows that, for this land-use variable, the majority of recorded percentages are within the range of 0–10% of the surface, while the remaining data values are scattered over the 10–100% interval. Similarly skewed distributions are found for all variables involved.

The data used for our study is composed of three data sets, one for each bird species of interest. Each data set includes a single discrete class variable that represents whether the bird species at hand is *present* or *absent* in a specific grid cell. The remaining variables, listed in Table 1, are continuous feature variables which represent the percentage (between 0% and 100%) of a grid cell's surface with a particular type of land use. The actual features were extracted from regional reports about the biology of each species, by selecting those pertaining to the species' habitat. Each data set contains 989 records, one per grid cell, and does not have any missing values, that is, for each grid cell, both the actual features and the associated class are recorded.

2.2. Classification models

Two types of Bayesian-network classifier are studied, each with discretized variables and with the original continuous variables respectively, and their performances are compared with those of a logistic-regression model.

2.2.1. Bayesian-network classifiers

A Bayesian network is a concise model of a joint probability distribution over a set of random variables (Jensen and Nielsen, 2007). It combines a directed acyclic graph, which describes the (in)dependencies between the variables, with local probability distributions per variable. From a Bayesian network, any probability of interest over its variables can be computed.

When used for classification purposes, a Bayesian network includes a designated class variable C. Of interest then is the posterior probability distribution $\Pr(C \mid \mathbf{f})$ over C given case observations $\mathbf{f}$ for the feature variables involved. To decide upon the class to which the observations $\mathbf{f}$ are to be assigned, two approaches are in use:

- *maximum-probability classification* (also known as "the winner takes all"), in which the case observations $\mathbf{f}$ are assigned to the most probable class given $\mathbf{f}$;
- *probability-threshold classification*, in which the observations $\mathbf{f}$ are assigned to the class c if $t_1 > \Pr(c \mid \mathbf{f}) \geq t_2$ for some suitable choice of decision thresholds t_1 and t_2.

For a binary class variable with the classes c_1 and c_2, probability-threshold classification with a decision threshold t assigns case observations $\mathbf{f}$ to c_1 if

$$Pr(c_1 \mid \mathbf{f}) \geq t$$

and to c_2 otherwise; taking $t = 0.5$ would then result in the same class assignment as maximum-probability classification. The overall performance of a probabilistic classifier is optimized by choosing a decision threshold based on the prior distribution over the class variable (Lachiche and Flach, 2003).

For classification purposes, tailored Bayesian networks with highly constrained graphical structures are in use, among which are the Naive Bayesian (NB) classifier and the Tree Augmented Naive Bayesian (TAN) classifier (Friedman et al., 1997). The Naive Bayesian classifier is the most constrained of all Bayesian-network classifiers: its graph consists of a designated node for the class variable and nodes modeling the feature variables with just this class variable for their parent. This type of classifier derives its name from the fact that its graphical structure captures the naive assumption that all feature variables are mutually independent given the class variable. Although this assumption does not generally hold in practice, NB classifiers tend to show quite competitive performance. TAN classifiers allow for explicitly representing dependencies among the feature variables by a tree structure, and in essence may thereby outperform NB classifiers.

[2] http://www.magrama.gob.es/es/biodiversidad/temas/inventarios-nacionales/inventario-especies-terrestres/.

[3] http://www.juntadeandalucia.es/medioambiente/site/rediam.

Table 1
Feature variables and prevalence (p) per species.

	Turdus viscivorus	*Cecropis daurica*	*Accipiter nisus*
	$p = 0.47$	$p = 0.84$	$p = 0.27$
1	Bare soil	Agricultural areas	Bare soil
2	Dams	Bare soil	Bare soil of scrub
3	Dense forest of conifers	Cliff	Dense forest of conifers
4	Dense forest of oaks	Dehesas	Dense forest of oaks
5	Dense scrubland	Dense forest	Dense grasslands
6	Dense scrubland of conifers	Dense scrubland	Dense scrubland
7	Dense scrubland of oaks	Dense scrubland of trees	Dense scrubland of oaks
8	Open scrubland	Open scrubland	Open grasslands
9	Grasslands of oaks	Open scrubland of trees	Open scrubland
10	Herbaceous crops	Grasslands	Open scrubland of oaks
11	Heterogeneous crops	Grasslands of trees	Grasslands of oaks
12	Irrigation pond	Herbaceous crops	Herbaceous crops
13	Olive crops	Heterogeneous crops	Heterogeneous crops
14	Other dense forests	Man-made water surfaces	Irrigated pool
15	Woody crops	River bed	Olive crops
16		Urban areas	Other disperse scrubland of trees
17		Woody crops	Other dense forest
18			Other dense scrubland of trees
19			River bed
20			Woody crops

Learning a Naive Bayesian classifier from a data set amounts to estimating probabilities from the available data so as to quantify the relationships between the class variable and each of the feature variables. Learning a TAN classifier in addition involves learning the graphical structure from the data. For this purpose, first a directed tree over the feature variables is learned by building upon the conditional mutual information between pairs of feature variables given the class variable (Chow and Liu, 1968); subsequently, the class variable is added and all modeled relationships are quantified.

2.2.2. Hybrid Bayesian networks

Bayesian networks were initially defined for discrete variables only. Even to date, Bayesian-network software packages tend to assume all variables to be discrete. As a consequence, upon developing a real-world application, all continuous domain variables have to be discretized by dividing their value ranges into a sequence of adjacent intervals. A probability distribution over the discretized variable then assigns to each such interval a single probability which can be viewed as approximating the continuous distribution over the interval by a constant function. In general, the use of more intervals upon discretization tends to result in a better approximation, albeit at the expense of a more complex model.

More recently, approaches have been developed that allow Bayesian networks to include both continuous and discrete variables (Langseth et al., 2012; Shenoy and West, 2011; Lauritzen and Jensen, 2001; Moral et al., 2001). In this paper we study Bayesian-network classifiers that employ *Mixtures of Truncated Exponentials* (MTEs) for their local probability distributions. Like discretization methods, MTE approaches divide the value range of a continuous variable into intervals. The continuous distribution per interval is then approximated by an exponential function rather than by a constant function (Rumí, 2003). Similar to discretization, the use of more intervals tends to result in a better approximation, but will also yield a more complex model. By including more terms in the MTE per interval, the approximation also tends to improve, yet again at the cost of a more complex model (Rumí and Salmerón, 2007; Morales et al., 2006; Rumí et al., 2006).

2.2.3. Logistic regression

Logistic regression is a type of regression in which a binary response variable (the binary class variable, in terms of our classification context) is related to multiple explanatory variables (the feature variables) which may be discrete or continuous (Scott, 2010). Upon classification of case observations $\mathbf{f}$ for the explanatory variables, the response with highest posterior odds given $\mathbf{f}$ is determined and assigned to the case. Logistic-regression classification thereby in essence is similar to taking a maximum-probability approach to classification. In fact, logistic-regression classification is known to be equivalent to Naive-Bayesian classification under mild conditions (Roos et al., 2005).

2.3. Data discretization

In our study, four discretization methods are compared: *Equal Frequency*, *Equal Width*, *Chi-Merge* and *Minimum Description Length Principle* (*MDLP*) discretization. These four methods are the most commonly studied discretization methods in the literature (García et al., 2013; Liu et al., 2002). In environmental modeling with Bayesian networks, the *Equal Frequency* and *Equal Width* methods prevail (Aguilera et al., 2011). The *Chi-Merge* method has, to the best of our knowledge, never been used in such applications, while use of the *Minimum Description Length Principle* has been reported in just a single environmental-modeling study (Fernandes et al., 2013).

2.3.1. Discretization methods

Application of a discretization method to a data set starts by sorting the available data points in increasing order of their value for the continuous variable to be discretized. The data points are then distributed over $k > 1$ bins, each of which is associated with an interval from the variable's overall value range. Discretization methods differ in whether or not the number of intervals k is chosen beforehand and in how the boundaries, or cut points, for the intervals are determined.

Equal Frequency (EF) and Equal Width (EW) discretisation The *Equal Frequency* and *Equal Width* methods are probably the simplest discretization methods in use (Liu et al., 2002). With the *Equal Frequency* method, each constructed interval includes essentially the same number of data points. With the *Equal Width* method, all constructed intervals have equal length; these intervals may thus have varying numbers of data points. With both *Equal Frequency* and *Equal Width*, the parameter k dictating the number of intervals used for the discretization, is chosen beforehand. Upon discretizing the continuous variables underlying a data set, in essence different k's may be chosen per variable. In most applications, however, a single k is used for all variables concerned. In environmental sciences, the number of intervals is often chosen based upon expert knowledge (Chen and Pollino, 2012); without such knowledge, an appropriate number may be found by experimentation. Alternatively, the number of intervals k may be decided upon by the *Proportional k-interval Discretization* (PKID) guideline (Yang and Webb, 2009), which takes $k = \sqrt{N}$ where N is the number of data points available.

The *Equal Frequency* and *Equal Width* methods are the most commonly used methods for discretizing continuous variables in environmental modeling with Bayesian networks (Aguilera et al., 2011). Chen and Pollino (2012) argue however, that the *Equal Width* method is less suited for data sets that have a markedly uneven distribution or include prominent outliers, and that the *Equal Frequency* method is less appropriate for data sets in which specific values are overrepresented.

Chi-Merge (ChiM) discretization Chi-Merge is a supervised discretization method which takes the classes associated with the available data points into account (Kerber, 1992). The method starts by constructing a sequence of intervals such that each interval includes a single data point. The χ^2-statistic is then used to decide whether two adjacent intervals be merged. For this purpose, for each pair of adjacent intervals, the χ^2-value is calculated from:

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{m}\frac{(A_{ij}-E_{ij})^2}{E_{ij}} \tag{1}$$

where m is the number of distinct classes, A_{ij} is the number of data points in interval i that are of class j, and E_{ij} is the expected number of data points of class j in interval i under the assumption that the class frequencies per interval are the same; this expected number E_{ij} is established from:

$$E_{ij} = \frac{\left(\sum_j \mathrm{A}_{ij}\right)\cdot\left(\sum_i \mathrm{A}_{ij}\right)}{\sum_i \sum_j \mathrm{A}_{ij}} \tag{2}$$

In each iteration of the *Chi-Merge* method, the pair of adjacent intervals with the smallest χ^2-value are merged, provided that this value falls below the confidence threshold $\chi^2_{df,\alpha}$ read from the χ^2-distribution table, where df is the number of degrees of freedom $m - 1$ and α is a user-specified significance level (preferably between 0.9 and 0.99). The iterative procedure halts when all χ^2-values are above the confidence threshold.

Since the *Chi-Merge* method serves to discretize each continuous variable in a data set independently, the number of intervals constructed per variable may differ. In order to avoid large numbers of intervals in practice, a maximum number of intervals can be pre-set upon application of the *Chi-Merge* method. The iterative procedure described above is then halted as this number of intervals is reached.

Minimum description length principle (MDLP) discretization Similar to *Chi-Merge*, the *MDLP* method, first introduced by Fayyad and Irani (1993, 1996), is a supervised discretization method which takes the classes associated with the data points into account. Starting with a single interval composed of all data points sorted in increasing order of their value for the variable to be discretized, *MDLP* constructs, by an iterative procedure, a sequence of intervals over the variable's overall value range. Within an interval S, potential cut points t_i are defined between each pair of data values; such a cut point would in essence partition the interval S into the two adjacent intervals S_1^i and S_2^i. For each potential cut point t_i in S, the *Class Information Entropy* (CIE) of the partition induced by t_i is then computed, from:

$$CIE(S, t_i) = \frac{N(S_1^i)}{N(S)}\cdot E(S_1^i) + \frac{N(S_2^i)}{N(S)}\cdot E(S_2^i) \tag{3}$$

where S_1^i, S_2^i are the two (sub)intervals that would be induced by the cut point t_i, $N(\cdot)$ denotes the number of data points in the indicated interval, and $E(\cdot)$ is the entropy of the class distribution estimated from the data points in the indicated interval. This entropy $E(S')$ for an interval S' is calculated as:

$$E(S') = -\sum_{c_j} P_{S'}(c_j)\cdot \log_2 P_{S'}(c_j) \tag{4}$$

where c_j is a class and $P_{S'}(c_j)$ is the estimated probability of occurrence of c_j in the interval S'. The potential cut point t_i with the smallest *Class Information Entropy* now is accepted as an actual cut point, provided that doing so yields an information gain $E(S) - CIE(S, t_i)$ satisfying the following criterion:

$$E(S) - CIE(S, t_i) > \frac{1}{N(S)}\cdot(\log_2(N(S) - 1) + \triangle(S, t_i)) \tag{5}$$

where $\triangle(S, t_i)$ equals

$$\triangle(S, t_i) = \log_2(3^m - 2) - [m\cdot E(S) - m_1\cdot E(S_1^i) - m_2\cdot E(S_2^i)] \tag{6}$$

with m, m_j, j = 1,2, being the number of distinct classes in the intervals S, S_j^i, respectively. This procedure is repeated iteratively, with the two intervals S_1^i and S_2^i substituted for the interval S for each accepted cut point t_i, as long as there is at least one potential cut point that satisfies the information-gain criterion above. After the procedure has halted, the accepted cut points serve to define the intervals from the overall value range of the variable being discretized.

MDLP discretization is one of the more commonly used discretization methods in general (García et al., 2013). Experiments by Liu et al. (2002) suggest that *MDLP* in fact is one of the best performing discretization methods in practice. Despite its reported good performance, however, *MDLP* has hardly been used in environmental modeling with Bayesian networks (Fernandes et al., 2013).

2.3.2. *The discretized data sets for the study*

The continuous variables of the species data sets described in Section 2.1, were discretized by the four methods reviewed above, as available from the Discretization Package[4] of the R statistical computing software. With each of the *Equal Frequency* and *Equal Width* methods, four different discretizations were constructed for the variables under study, with 3, 5, 10 and 32 intervals, respectively, where the PKID criterium gave rise to the number of intervals $k = \sqrt{N} = \sqrt{989} = 32$. For application of the *Chi-Merge* method, a significance level of 0.99 was used as suggested in the literature; no limit was set on the number of intervals.

Discretization resulted in 10 data sets per species: four data sets resulted from using the *Equal Frequency* method with 3, 5, 10 and 32 intervals, respectively, and four resulted from using the *Equal Width* method with the same numbers of intervals; one data set resulted from application of the *Chi-Merge* method, and one set of discretized data was constructed using *MDLP*. As per species moreover, the original continuous data were used in the investigations, our study involved a total of 33 data sets.

2.4. *Model learning and validation*

From each of the 30 discretized data sets, discrete Naive Bayesian and TAN classifiers were learned. From the three continuous data sets, we constructed NB and TAN classifiers with Mixtures of Truncated Exponentials for the local probability distributions; the number of exponentials was set to three based on preliminary experimentation. All classifiers were learned using the

[4] https://cran.r-project.org/web/packages/discretization/index.html.

Table 2
Minimum, maximum and mean number of intervals constructed by the *Chi-Merge* and *MDLP* discretization methods, per species data set.

		Cecropis daurica	*Turdus viscivorus*	*Accipiter nisus*
Chi-Merge	Mean	52.9	12	18.7
	Minimum	14	6	9
	Maximum	91	25	29
MDLP	Mean	1.5	2.4	1.8
	Minimum	1	1	1
	Maximum	2	4	3

Elvira software[5] (Elvira-Consortium, 2002). In addition, 33 logistic-regression models were constructed using the R statistical software package.

To arrive at reliable estimates of the predictive performance per model, a *ten-fold cross validation* procedure was used. To this end, each data set D was partitioned into ten equally-sized disjoint subsets, or *folds*, D_i, $i = 1, \ldots, 10$. Then, for each fold, the following procedure was run:

- set the current fold D_i aside for testing;
- learn the appropriate type of classification model from the set $D^{-i} = \bigcup_{j=1,\ldots,10, j \neq i} D_j$ composed of the data from the other nine folds;
- estimate the performance of the thus learned model by classifying the data points from D_i.

The predictive performance of the classifier learned from the entire data set D now is estimated as the performance result averaged over the ten runs.

As measures of performance for the learned classification models, the well-known *sensitivity* and *specificity* characteristics are used. Estimates for these characteristics are calculated from:

$$sensitivity = \frac{TP}{TP + FN} \tag{7}$$

$$specificity = \frac{TN}{FP + TN} \tag{8}$$

where *TP* is the number of *true positives*, that is, the number of data points in the test set in which the species is known to be *present* and which are assigned to the *present* class by the learned classifier, and *TN* is the number of *true negatives*, that is, the number of data points in the test set in which the species is *absent* and which are assigned to the *absent* class by the classifier; *FP* is the number of *false positives*, that is, the number of data points in the test set which are classified as *present*, yet are known to be *absent*, and *FN* is the number of *false negatives*, that is, the number of data points which are classified as *absent* and are known to be *present*. The thus obtained *sensitivity* and *specificity* estimates are combined into an *averaged performance* estimate through

$$Averaged \quad performance = \frac{1}{2}(sensitivity + specificity) \tag{9}$$

Since the *sensitivity* and *specificity* estimates found for a classifier are dependent of the decision threshold used upon classification, all Bayesian-network classifiers were validated using maximum-probability classification to allow a fair comparison with their matching logistic-regression models. The Bayesian-network classifiers were also validated using probability-threshold classification with the prevalences of the various bird species as the decision thresholds. All in all, with each of the 33 data sets, five classification models were learned and validated: an NB, a TAN and a logistic-regression model were constructed and evaluated using maximum-probability classification, and an NB and a TAN were learned and validated using probability-threshold classification.

3. Results

The experimental results from using the different discretization methods on the various data sets are summarized by the granularity of the resulting discretizations and by the performance of the learned classification models.

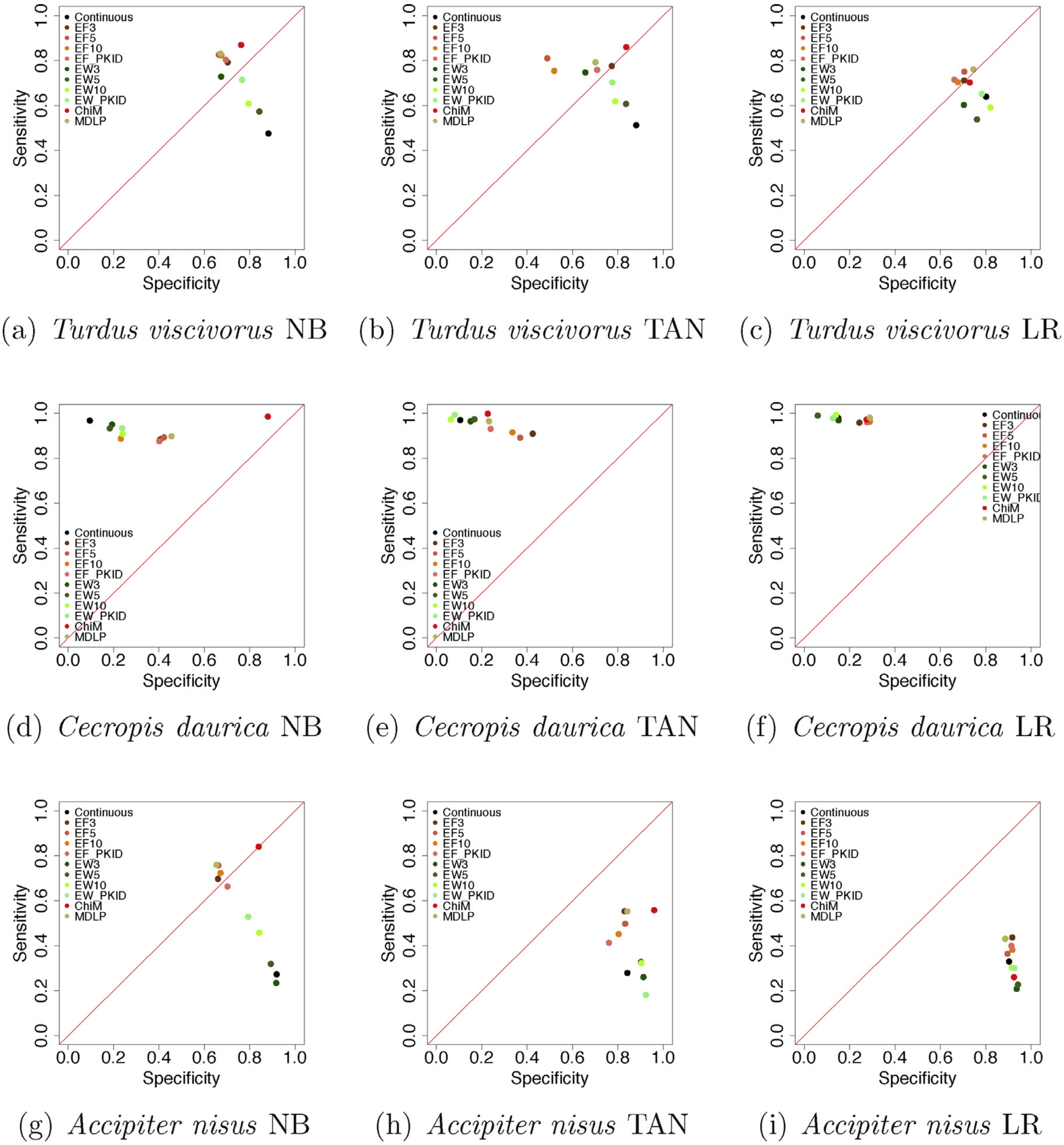
3.1. Granularity of discretization

For use of the *Equal Frequency* and *Equal Width* discretization methods, the numbers of intervals to be constructed for a continuous variable were chosen beforehand, as 3, 5, 10 and 32, respectively; for each variable, the same number of intervals was used. With the *Chi-Merge* and *MDLP* methods, the numbers of intervals to be constructed were not pre-set but rather established by the methods themselves, for each variable separately. Table 2 reports, for each species data set, the numbers of intervals constructed by the *Chi-Merge* and *MDLP* methods, respectively; the means reported in the table were calculated by averaging over the discretizations of all variables in the data set at hand. The table shows that, for our data sets, the *MDLP* method resulted in quite coarse discretizations, with just a limited number of intervals per variable. The *Chi-Merge* method, on the other hand, resulted in more fine-grained discretizations, with over 50 intervals for some of the variables involved.

Table 3
Averaged performance estimates of the classification models with maximum-probability classification, per species data set.

	Turdus viscivorus			*Cecropis daurica*			*Accipiter nisus*		
	NB	TAN	LR	NB	TAN	LR	NB	TAN	LR
Continuous	0.68	0.69	0.72	0.53	0.54	0.57	0.60	0.54	0.59
EF3	0.75	0.75	0.71	0.64	0.67	0.60	0.68	0.69	0.68
EF5	0.75	0.74	0.73	0.66	0.64	0.63	0.71	0.67	0.63
EF10	0.75	0.75	0.69	0.66	0.63	0.62	0.70	0.63	0.65
EF_PKID	0.75	0.68	0.69	0.64	0.59	0.62	0.68	0.59	0.66
EW3	0.70	0.71	0.65	0.57	0.56	0.57	0.58	0.59	0.57
EW5	0.70	0.73	0.65	0.56	0.57	0.53	0.61	0.62	0.58
EW10	0.70	0.70	0.71	0.58	0.52	0.57	0.65	0.61	0.61
EW_PKID	0.74	0.74	0.72	0.59	0.54	0.55	0.66	0.55	0.62
ChiM	0.82	0.82	0.72	0.93	0.61	0.63	0.84	0.76	0.59
MDLP	0.75	0.78	0.75	0.68	0.60	0.63	0.71	0.70	0.66

[5] Elvira is a public-domain Java-based software package: http://leo.ugr.es/elvira.



(a) *Turdus viscivorus* NB (b) *Turdus viscivorus* TAN (c) *Turdus viscivorus* LR

(d) *Cecropis daurica* NB (e) *Cecropis daurica* TAN (f) *Cecropis daurica* LR

(g) *Accipiter nisus* NB (h) *Accipiter nisus* TAN (i) *Accipiter nisus* LR

Fig. 3. *Sensitivity* and *specificity* estimates for the NB, TAN and logistic-regression (LR) models with maximum-probability classification, per species data set.

3.2. Maximum-probability classification

From each of the species data sets, Bayesian-network classifiers and logistic-regression models were learned as described in Section 2.4. The performances of the learned models using maximum-probability classification are visualized in Fig. 3, which shows the *sensitivity* and *specificity* estimates found; Table 3 summarizes these estimates in the models' *averaged performances*.

For the data set pertaining to *Turdus viscivorus*, all logistic-regression models showed quite similar performance, regardless of whether continuous or discretized data were used and, in the latter case, regardless of the discretization method employed (Fig. 3(c)). The Bayesian-network classifiers (Fig. 3(a) and (b)) showed more divergence in their performance characteristics. From among the discretization methods used, the *Chi-Merge* method resulted in the best balance of the *specificity* and *sensitivity* characteristics estimated for the classifiers, with an *averaged performance* of 0.82. Fig. 3(a) and (b) further show that the continuous Bayesian-network classifiers had a worse *sensitivity* than the classifiers learned from discretized data.

While the data set pertaining to *Turdus viscivorus* is well balanced with respect to the two classes, the other two data sets are less balanced, with a prevalence of 84% for the *Cecropis daurica* and a prevalence of 27% for the *Accipiter nisus*, respectively. For these less balanced data sets, all constructed models were found to excel at predicting the most probable class. More specifically, for the *Cecropis daurica* all classifiers attained a high *sensitivity* (Fig. 3(d)–(f)), while for the *Accipiter nisus* the classifiers attained a high *specificity* (Fig. 3(g)–(i)).

For the data set pertaining to *Cecropis daurica*, all Bayesian-network classifiers showed quite similar performance, yet with a notable single exception. The Naive Bayesian classifier learned from the data after discretization with the *Chi-Merge* method, showed very good performance in terms of both *sensitivity* and *specificity*; this classifier in fact resulted in an *averaged performance* of 0.93 (Table 3). Also for the *Accipiter nisus* data set discretized with the

Table 4
Averaged performance estimates of the classifiers using probability-threshold classification, per species data set.

	Turdus viscivorus		*Cecropis daurica*		*Accipiter nisus*	
	NB	TAN	NB	TAN	NB	TAN
Continuous	0.68	0.69	0.61	0.61	0.60	0.54
EF3	0.75	0.75	0.72	0.72	0.68	0.69
EF5	0.75	0.74	0.73	0.69	0.71	0.67
EF10	0.75	0.75	0.73	0.69	0.70	0.63
EF_PKID	0.75	0.68	0.65	0.61	0.68	0.59
EW3	0.70	0.71	0.63	0.62	0.58	0.59
EW5	0.70	0.73	0.63	0.63	0.61	0.62
EW10	0.70	0.70	0.64	0.61	0.65	0.61
EW_PKID	0.74	0.74	0.65	0.59	0.66	0.55
ChiM	0.82	0.82	0.96	0.70	0.84	0.76
MDLP	0.75	0.78	0.74	0.75	0.71	0.70

Chi-Merge method, did the NB classifier show the best balance of the *sensitivity* and *specificity* estimates attained, with an *averaged performance* of 0.84. While the TAN classifier never reached a *sensitivity* higher than 0.6 for the *Accipiter nisus* data set, the NB classifier gave *sensitivity* estimates higher than 0.7 after discretizing the data with the *Equal Frequency* method with 3, 5 and 10 intervals, with *MDLP* and with the *Chi-Merge* method.

The performance characteristics of the logistic-regression models learned from the *Cecropis daurica* and *Accipiter nisus* data sets (Fig. 3(f) and (i)) again were hardly affected by the discretization method used. For these two data sets, the *averaged performance* estimates found with the logistic-regression models were in the 0.53–0.68 range (Table 3). While for all models very good performance at predicting the most probable class was seen, the best *specificity* achieved by these models for *Cecropis daurica* was smaller than 0.3; similarly, the best *sensitivity* achieved for *Accipiter nisus* by the logistic-regression models was below 0.45.

3.3. *Probability-threshold classification*

The performances of the Bayesian-network classifiers learned from each of the species data sets were once more investigated, this time using probability-threshold classification. The detailed results are visualized in Fig. 4, in terms of the *sensitivity* and *specificity* estimates found; Table 4 summarizes these estimates in the models' *averaged performances*.

For the data set pertaining to *Turdus viscivorus*, the learned Bayesian-network classifiers were found to exhibit similar performance with probability-threshold classification as with maximum-probability classification (Fig. 4(a) and (b)). Since the *Turdus viscivorus* data set includes a binary class variable, maximum-probability classification was equivalent to probability-threshold classification with a decision threshold equal to 0.5. Based on the prevalence for *Turdus viscivorus*, probability-threshold classification was performed with a threshold of 0.47. Given the small difference in decision threshold used, similar performance of the Bayesian-network classifiers under the two types of classification was not unexpected.

Also for the data set pertaining to *Accipiter nisus* were the performances of the Bayesian-network classifiers with probability-threshold classification comparable to those found with maximum-probability classification (Fig. 4(e) and (f)). With this data set, however, the decision threshold for classification was set to the prevalence 0.27 of the bird species, which differed substantially from the 0.5 threshold used with maximum-probability classification.

When validated on the data set pertaining to *Cecropis daurica*, the Bayesian-network classifiers showed a different performance with probability-threshold classification (Fig. 4(c) and (d)) than with maximum-probability classification. In fact, use of the species' prevalence of 0.84 as the decision threshold for classification resulted in a better balance of the *sensitivity* and *specificity* characteristics estimated for the classifiers (Table 4) than use of the 0.5 threshold with maximum-probability classification. For the Naive Bayesian classifiers specifically, the good performances in terms of *sensitivity* were matched by a *specificity* between 0.7 and 0.8 after discretizing the data with *MDLP* and with the *Equal Frequency* method with three intervals; the corresponding *specificity* estimates found with maximum-probability classification were below 0.4. For the NB classifier moreover, discretization of the data with the *Chi-Merge* method gave the best *averaged performance* estimate, equal to 0.96. For the TAN classifier, discretization with the *MDLP* method gave the best result.

From the detailed *sensitivity* and *specificity* estimates plotted for the various Bayesian-network classifiers in Figs. 3 and 4, a general pattern emerges. Both with maximum-probability classification and with probability-threshold classification, discretization of the data with the *Equal Width* method tends to result in classifiers with a good performance at predicting the most probable class, that is, the *C. daurica* being *present* and the *A. nisus* being *absent*. With both types of classification, moreover, discretization with the *Equal Frequency* method tends to result in a better balance of the *sensitivity* and *specificity* characteristics of the learned Bayesian-network classifiers.

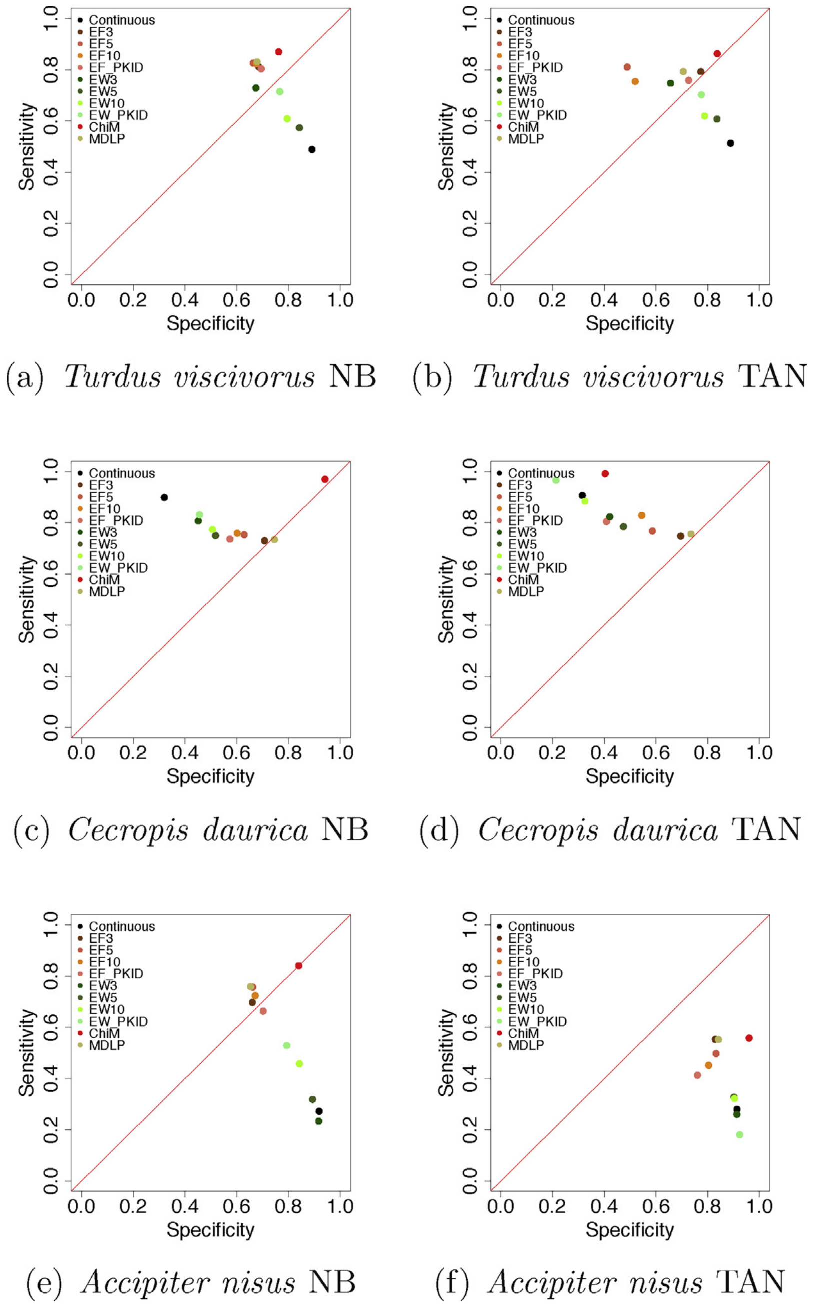
4. Discussion

Based on the experimental results described in Section 3, we discuss some implications for use of the various discretization methods and classification models in species distribution modeling.

Logistic-regression models Regression methods are widely used in environmental modeling in general (Schmitz et al., 2005) and for species distribution modeling in particular (Li and Wang, 2013). For well-balanced data sets, in which a species is (more or less) equally likely to be *present* as it is to be *absent*, our experimental results suggest that logistic-regression models can attain relatively high *sensitivity* and *specificity* characteristics. The overall performance of these models moreover, appears not to be affected by discretization of the data nor by the method used if the data were discretized. For less balanced data sets, however, logistic-regression models tend to fail at predicting the least probable class.

From an environmental point of view, a species distribution model should accurately predict the presence of a specific species in a territory, that is, it should show a high *sensitivity*. For abundant species, such as *Cecropis daurica* in our study, logistic-regression models can indeed attain a high *sensitivity* and thereby show satisfactory performance. In real-world applications, however, attention will mostly focus on endangered or rare species, such as *Accipiter nisus*. Our experimental results suggest that, for such species, logistic-regression models may not be able to achieve a satisfactory performance. For such species, therefore, using logistic regression may not be the best possible choice.

Bayesian-network classifiers and the effect of decision thresholds One of the advantages of Bayesian-network classifiers over other types of classifier is that the classification decision is separated from the prediction process (Uusitalo, 2007). The Bayesian networks underlying these classifiers in essence return a posterior probability distribution over the class variable given the case observations, based upon which a classification decision is taken. As discussed in



(a) *Turdus viscivorus* NB (b) *Turdus viscivorus* TAN

(c) *Cecropis daurica* NB (d) *Cecropis daurica* TAN

(e) *Accipiter nisus* NB (f) *Accipiter nisus* TAN

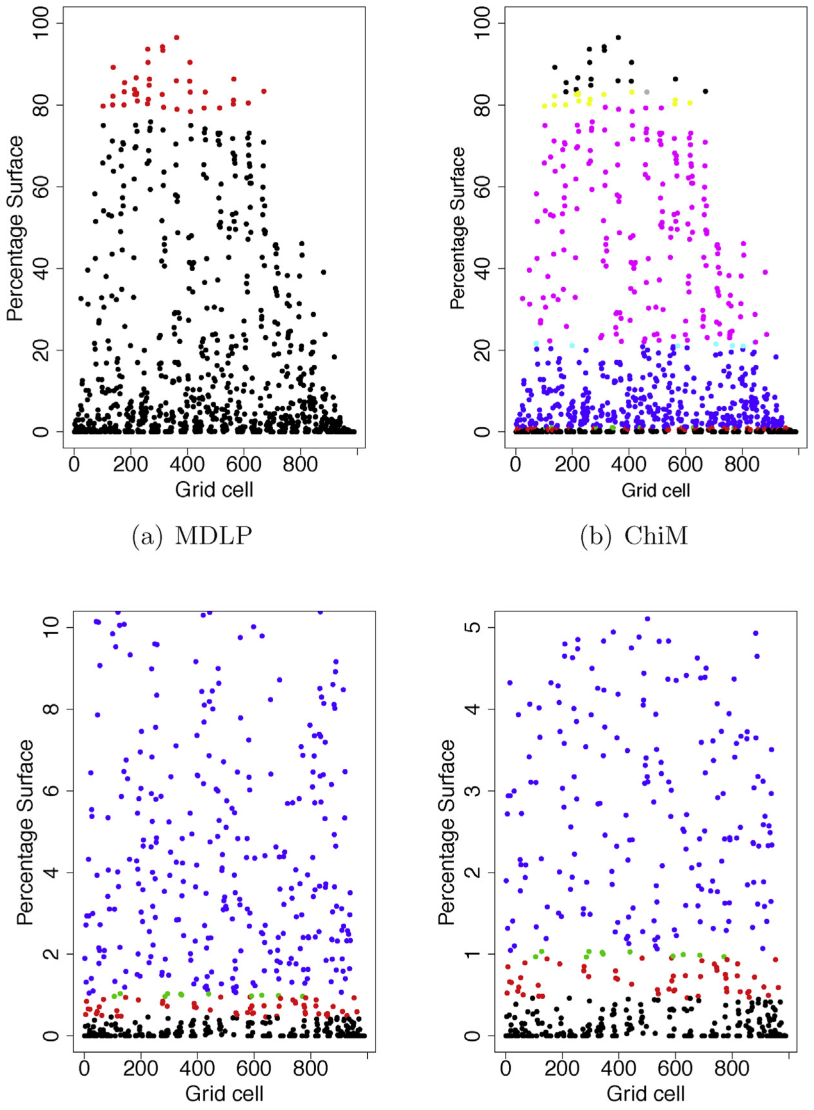
Fig. 4. *Sensitivity* and *specificity* estimates for the NB and TAN classifiers with probability-threshold classification, per species data set.

the previous sections, cases can then be assigned to the most probable class or to a class determined through a probability threshold.

Naive Bayesian and TAN classifiers are known to show a tendency to produce rather skewed posterior distributions for their class variable (Bennett, 2000). As a consequence, choosing non-extreme thresholds with probability-threshold classification may not dramatically change performance compared to maximum-probability classification. For the *Turdus viscivorus* and *Accipiter nisus* data sets in our study, in fact, classification with the decision thresholds of 0.47 and 0.27, respectively, did not result in a performance different from using the 0.5 threshold of maximum-probability classification. For the former species, this experimental finding was not unexpected given the small difference between the thresholds of 0.47 and 0.5. For the latter species, the difference between the two thresholds involved was more substantial. The finding of similar performance with the two types of classification now indicates that, for none or just a few cases, the established posterior probability of the species being *present* was in the 0.27–0.5 range, which would indeed be explained by the tendency of the Bayesian-network classifiers to produce rather skewed distributions over their class variable. While for *Turdus viscivorus* and *Accipiter nisus* using the species' prevalence for the decision threshold did not have any impact on classification performance, for the *Cecropis daurica* species the performance of the Bayesian-network classifiers did improve with probability-threshold classification using the more extreme decision threshold probability of 0.84.

The above insights from our experimental results suggest that using probability-threshold classification can be beneficial with Bayesian-network classifiers developed for species with quite small prevalences.

Continuous Bayesian-network classifiers Direct use of available continuous data is often recommended for Bayesian-network



(a) MDLP

(b) ChiM

(c) ChiM (0-10% and 0 - 5% of surface)

Fig. 5. Distribution of the *Olive croplands* variable, discretized by the *MDLP* method (a) and by the *Chi-Merge* method (b), with a detailed view of the latter discretization for 0–10% and 0–5%, respectively, of the surface (c).

learning, to avoid loss of information due to discretization (Uusitalo, 2007). The current generation of Bayesian networks can cope with continuous probability distributions only to some extent, however: local distributions for the continuous variables are required to be Gaussian (Lauritzen and Wermuth, 1989) or are approximated by polynomial or exponential functions, such as the MTEs used in our study.

In our experimental study, the Bayesian-network classifiers with MTEs for their local distributions showed good performance at predicting the most probable class. Since typically a large number of data points is required to allow satisfactory approximation of the continuous distributions at hand, this good performance may be attributed, to at least some extent, to the availability of many data points from the predominant class. For endangered or rare species, where the class of interest is the less probable one, our experimental results suggest that Bayesian-network classifiers with MTEs may result in relatively poor *sensitivity* and hence show unsatisfactory performance. For such species, direct use of available continuous data may not be the best choice for finding Bayesian-network classifiers of good performance.

Unsupervised discretization The unsupervised *Equal Width* and *Equal Frequency* discretization methods are widely used in environmental modeling through Bayesian networks (Aguilera et al., 2011). Chen and Pollino (2012) already argued that both methods are suitable for discretizing variables with a more or less even distribution over their values. They further argued that use of the *Equal Width* method is less appropriate for data sets that have a markedly uneven distribution or include prominent outliers, and that the *Equal Frequency* method is less suited for data sets in which specific values are overrepresented. The land-use variables in our study typically do not have even distributions, as was illustrated for the *Olive cropland* variable in Fig. 2(b).

The *Equal Frequency* method partitions the overall value range of a continuous variable into *k* intervals such each interval includes an essentially equal number of data points. Yet, data points with the same value for the continuous variable to be discretized are never placed in different intervals. Since the feature variables in our study capture types of land use that are present in relatively few grid cells, for any such variable a large number of data points include the value 0. These data points are all included in the first interval, therefore, and the remaining data points are equally distributed over the remaining $k-1$ intervals.

Our experimental results indicate that, with *Equal Frequency* discretization, all constructed Bayesian-network classifiers show good performance at predicting the most probable class; this good performance is generally balanced by a reasonable performance for the less probable class. The results further show that using just three intervals for the discretization tends to result in the best balance of *sensitivity* and *specificity* characteristics for all Bayesian-network classifiers learned. Since for most land-use variables a large number of data points include the value 0, the majority of the intervals constructed with the *Equal Frequency* method will include just a few data points. In fact, the more intervals are used, the fewer data points are expected per interval and the less informative the intervals tend to become for classification purposes. Using a small number of intervals therefore appears to be the best option upon *Equal Frequency* discretization of data sets in which specific values are overrepresented.

The *Equal Width* discretization method partitions the overall value range of a continuous variable into k intervals such that all intervals are of equal length. Just like the *Equal Frequency* method, it includes all data points with the value 0 for the variable to be discretized in the first interval. Since for most land-use variables a large number of data points include this value and the method aims at constructing intervals of equal length, actually the majority of data points will be included in this first interval and very few points remain for the subsequent intervals, which causes these intervals to be rather uninformative.

Our experimental results now indicate that, with *Equal Width* discretization, all constructed Bayesian-network classifiers show good performance at predicting the most probable class, just as with *Equal Frequency* discretization. With *Equal Width* discretization, however, this good performance is balanced by a relatively poor performance for the less probable class, as a consequence of the constructed highly dominant first interval. While, with *Equal Frequency* discretization, using three intervals resulted in the best balance of *sensitivity* and *specificity* characteristics for all Bayesian-network classifiers learned, the experimental results from *Equal Width* discretization suggest that using a small number of intervals may not always give a well-balanced performance. For the *Acciptiter nisus* data set, with the low prevalence of its species, in fact, using three intervals for the discretization resulted in a very high specificity while more intervals were required to attain a reasonable sensitivity.

Supervised discretization The supervised *Chi-Merge* and *MDLP* methods take the classes associated with the available data points into account upon discretizing the continuous variables involved. The two methods differ in their starting points for the iterative procedure and in their criteria for merging and splitting intervals. The *Chi-Merge* method starts with a separate interval per data point and iteratively merges two adjacent intervals if the class distributions in these intervals are more or less similar. The *MDLP* method on the other hand, starts with a single interval including all data points and iteratively splits an interval if the class distributions in the resulting subintervals are more skewed than the distribution in the original interval.

As argued above, the continuous land-use variables in our study have highly skewed distributions, as a result of the heterogeneous conditions of Andalusia. From among the two supervised discretization methods reviewed in our study, the *Chi-Merge* method seems better able to capture the characteristics of the data than the *MDLP* method. As an example, Fig. 5 depicts the available data points pertaining to *Olive croplands* coverage of the cells of the UTM grid. The full range of the percentage of surface covered, is partitioned into intervals by the two discretization methods. The discretization constructed by the *MDLP* method is shown in Fig. 5(a) and the discretization by *Chi-Merge* is shown in Fig. 5(b); the various partitions are indicated in color. Fig. 5(a) reveals that, with the *MDLP* method, just two intervals were constructed for the entire range of the percentage of covered surface. With the *Chi-Merge* method, multiple intervals were created: four intervals were constructed for the lower percentages of surface coverage (Fig. 5(c)) and four more intervals resulted for the rest of the percentage range. The difference between the resulting discretizations may, to some extent, be due to the stopping criteria employed by the two methods. Yet also the tendency of the *MDLP* method to construct intervals with class distributions of low entropy may cause this method to be less sensitive to small shifts in already quite skewed distributions than the *Chi-Merge* method is.

For all species data sets discretized with the *Chi-Merge* method, the Naive Bayesian classifiers attained high *sensitivity* and *specificity* characteristics, with averaged performances between 0.82 and 0.96. A similar trend was seen for the TAN classifiers constructed from the *Turdus viscivorus* data set discretized with *Chi-Merge*. For the less balanced data sets discretized with the *Chi-Merge* method, the TAN classifiers excelled at predicting the most probable class. For the least probable class, however, TAN classifiers with a better performance resulted from discretizing the data with less sophisticated methods. For the *Cecropis daurica* data set in fact, using the *Chi-Merge* method for discretization resulted in TANs with averaged performances of 0.61 and 0.70, while the best performing TANs had averaged performances of 0.67 and 0.72, respectively. The lesser performance found from using the *Chi-Merge* method with the *Cecropis daurica* data set may be attributed to the relatively large number of intervals constructed for the various feature variables: some of these intervals are likely to include only very few data points and, as a consequence, the strengths estimated for the dependencies involved in the TANs will most likely be unreliable.

For the species data sets discretized with the *MDLP* method, the performance trends of all Bayesian-network classifiers were more or less similar to those found for the sets discretized with the *Chi-Merge* method, although less prominent. Overall, the averaged performances of the various classifiers were found to lie below those of the corresponding classifiers learned from the *Chi-Merge* discretized data.

5. Conclusion and future research

In our experimental study, we compared the performances of different types of classification model and different discretization methods in view of species distribution modeling. In the study, we focused on prediction of the presence of various bird species in Andalusia from land-use data, and considered to this end three species with different prevalence rates. The experimental results obtained suggest that Bayesian-network classifiers, and among these especially the Naive Bayesian classifiers, may be preferable to logistic-regression models for the environmental-science context at hand. Our results further indicate that the *Chi-Merge* method may be the preferred method for discretizing the continuous variables involved, since with this method the best averaged performance results in terms of both *sensitivity* and *specificity* were found. As it is a supervised method, it is computationally more involved than the better known *Equal Frequency* and *Equal Width* methods for discretization. Implementations of the *Chi-Merge* method are available in software packages such as R for ready use in practice.

While most applications of Bayesian networks require discretization of the continuous variables underlying available data, only a restricted set of methods are used in practice. For species distribution modeling through Bayesian networks more specifically, further research efforts are required to gain insight in the foundational properties of the various discretization methods proposed in the literature and to establish their practical properties upon application to different types of environmental data. While the conclusions obtained from our experimental study are likely to hold for data sets with similar characteristics as our land-use data, the results cannot be directly extrapolated to other environmental data, such water quality, air pollution and climatic data, without further study. Moreover, since expert knowledge is often taken as a primary source of information in environmental science (Henriksen et al., 2007), and is in fact used for choosing cut points for discretization, the quality of expert-based discretizations should be compared with the discretizations found with automated methods. From a wider future perspective, it is worthwhile to study the strengths and weaknesses of using Bayesian networks for species distribution modeling compared to using the more common domain-specific models proposed in the literature, such as BIOCLIM and FLORAMAP.

Acknowledgements

R.F. Ropero is supported by the FPU research grant, AP2012-2117, funded by the Spanish Ministry of Education, Culture and Sport. The research presented in this paper was performed while R.F. Ropero stayed at Utrecht University; this research stay was supported by the Spanish Ministry of Education, Culture and Sport (grant EST15/00067).

References

Aguilera, P.A., Fernández, A., Fernández, R., Rumí, R., Salmerón, A., 2011. Bayesian networks in environmental modelling. Environ. Model. Softw. 26, 1376–1388.

Aguilera, P.A., Fernández, A., Ropero, R.F., Molina, L., 2013. Groundwater quality assessment using data clustering based on hybrid Bayesian networks. Stoch. Environ. Res. Risk Assess. 27 (2), 435–447.

Baur, B., Bozdag, S., 2015. A canonical correlation analysis-based dynamic Bayesian network prior to infer gene regulatory networks from multiple types of biological data. J. Comput. Biol. 22, 289–299.

Bennett, P., 2000. Assessing the calibration of Naive Bayes' posterior estimates. Tech. Rep. CMU-CS00-155. Carnegie Mellon University.

Busby, J., 1986. Bioclimate Prediction System (BIOCLIM). Users Manual Version 2.0. Australian Biological Resources, Study Leaflet, Canberra, Australia.

Chen, S.H., Pollino, C.A., 2012. Good practice in Bayesian network modelling. Environ. Model. Softw. 37, 134–145.

Chow, C.K., Liu, C.N., 1968. Approximating discrete probability distributions with dependence trees. IEEE Trans. Inf. Theory 14, 462–467.

Davison, A.C., Ramesh, N., 1996. Some models for discretized series of events. J. Am. Stat. Assoc. 91, 601–609.

Dedecker, A.P., Goethals, P.L.M., Gabriels, W., De Pauw, N., 2004. Optimization of Artificial Neural Network (ANN) model design for prediction of macroinvertebrates communities in the Zwalm river basin (Flanders, Belgium). Ecol. Model. 174, 161–173.

Dyer, F., ElSawah, S., Croke, B., Griffiths, R., Harrison, E., Lucena-Moya, P., Jakeman, A.J., 2014. The effects of climate change on ecologically-relevant flow regime and water quality attributes. Stoch. Environ. Res. Risk Assess. 28, 67–82.

Elith, J., Graham, C.H., Anderson, R.P., Dudik, M., Ferrier, S., Guisan, A., Hijmans, R.J., Huettmann, F., Leathwick, J.R., Li, J., Lohmann, L.G., Loiselle, B.A., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., Overton, J.M., Peterson, A.T., Phillips, S.J., Richardson, K.S., Scachetti-Pereria, S., Schapire, R.E., Soberón, J., Williams, S., Wisz, M.S., Zimmermann, N.E., 2006. Novel methods to improve prediction of species' distribution from occurrence data. Ecography 29, 129–151.

Elvira-Consortium, 2002. Elvira: an environment for creating and using probabilistic graphical models. Proceedings of the First European Workshop on Probabilistic Graphical Models, 222–230 http://leo.ugr.es/elvira.

Fayyad, U., Irani, K., 1993. Multi-interval discretization of continuous-valued attributes for classification learning. In: Thirteenth International Joint Conference on Artificial Intelligence. Morgan Kaufmann, San Mateo, CA, pp. 1022–1027.

Fayyad, U., Irani, K., 1996. Discretizing continuous attributes while learning Bayesian networks. In: Thirteenth International Conference on Machine Learning. Morgan Kaufmann, pp. 157–165.

Fernandes, J.A., Lozano, J.A., Inza, I., Irigoien, X., Pérez, A., Rodríguez, J.D., 2013. Supervised pre-processing approaches in multiple class variables classification for fish recruitment forecasting. Environ. Model. Softw. 40, 245–254.

Friedman, N., Geiger, D., Goldszmidt, M., 1997. Bayesian network classifiers. Mach. Learn. 29, 131–163.

Fukuda, S., De Baets, B., Waegeman, W., Verwaeren, J., Mouton, A., 2013. Habitat prediction and knowledge extraction for spawning European grayling (*Thymallus thymallus* L.) using a broad range of species distribution models. Environ. Model. Softw. 47, 1–6.

García, S., Luengo, J., Saez, J.A., Lopez, V., Herrera, F., 2013. A survey of discretization techniques: taxonomy and empirical analysis in supervised learning. IEEE Transactions on Knowledge and Data Engineering, 734–750.

Henriksen, H.J., Rasmussen, P., Brandt, G., von Bülow, D., Jensen, F.V., 2007. Public participation modelling using Bayesian networks in management of groundwater contamination. Environ. Model. Softw. 22, 1101–1113.

Jensen, F.V., Nielsen, T.D., 2007. Bayesian Networks and Decision Graphs. Springer.

Jones, P., Gladkov, A., 1999. FloraMap: A Computer Tool for Predicting the Distribution of Plants and Other Organisms in the Wild. Centro Internacional de Agricultura Tropical (CIAT), Cali, Colombia.

Kerber, R., 1992. Chimerge: discretization of numeric attributes. In: AAAI-92, Ninth National Conference Artificial Intelligence. AAAI Press/The MIT Press, pp. 123–128.

Lachiche, N., Flach, P., 2003. Improving accuracy and cost of two-class and multi-class probabilistic classifiers using ROC curves. In: Fawcett, T., Mishra, N. (Eds.), Proceedings of the Twentieth International Conference on Machine Learning. AAAI Press, Menlo Park, pp. 416–423.

Langseth, H., Nielsen, T.D., Rumí, R., Salmerón, A., 2012. Mixtures of truncated basis functions. Int. J. Approx. Reason. 53 (2), 212–227.

Lauritzen, S.L., Jensen, F., 2001. Stable local computation with conditional Gaussian distributions. Statistics and Computing, Vol. 11., pp. 191–203.

Lauritzen, S.L., Wermuth, N., 1989. Graphical models for associations between variables, some of which are qualitative and some quantitative. Ann. Stat. 17, 31–57.

Li, X., Wang, Y., 2013. Applying various algorithms for species distribution modelling. Integr. Zool. 8 (2), 124–135.

Li, Y., 2007. Control of spatial discretisation in coastal oil spill modelling. Int. J. Appl. Earth Observ. 9, 392–402.

Lima, M.D., Nassar, S.M., Rodriges, P.I.R., Freitas-Filho, P., Jacinto, C.M.C., 2014. Heuristic discretization method for Bayesian networks. J. Comput. Sci. 10 (5), 869–878.

Liu, H., Hussain, F., Lim, C., Dash, M., 2002. Discretization: an enabling technique. Data Mining Knowl. Discov. 6, 393–423.

Liu, Y., Zhang, W., Zhang, Z., 2015. A conceptual data model coupling with physically based distributed hydrological models based on catchment discretization schemas. J. Hydrol. 530, 206–215.

Maldonado, A., Ropero, R.F., Aguilera, P., Rumí, R., Salmerón, A., 2015. Continuous Bayesian networks for the estimation of species richness. Prog. Artif. Intell. 4, 49–57.

Moral, S., Rumí, R., Salmerón, A., 2001. Mixtures of Truncated Exponentials in hybrid Bayesian networks. In: ECSQARU'01. Lecture Notes in Artificial Intelligence, vol. 2143. Springer, pp. 156–167.

Morales, M., Rodríguez, C., Salmerón, A., 2006. Selective Naïve Bayes predictor using mixtures of truncated exponentials. Proceedings of the International Conference on Mathematical and Statistical Modelling (ICMSM'06).

Myers, N., Mittenmeier, R.A., Mittenmeier, C.G., da Fonseca, G.A.B., Kent, J., 2000. Biodiversity hotspots for conservation priorities. Nature 403, 853–858.

Nash, D., Waters, D., Buldu, A., Wu, Y., Lin, Y., Yang, W., Song, Y., Shu, J., Qin, W., Hannah, M., 2013. Using a conceptual Bayesian network to investigate environmental management in vegetable production in the Lake Taihu region of China. Environ. Model. Softw. 46, 170–181.

Newton, A.C., Stewart, G.B., Díaz, A., Golicher, D., Pullin, A.S., 2007. Bayesian Belief Networks as a tool for evidence-based conservation management. J. Nat. Conserv. 15, 144–160.

Park, M.H., Stenstrom, M.K., 2008. Classifying environmentally significant urban land uses with satellite imagery. J. Environ. Manag. 86, 181–192.

Pollino, C.A., White, A.K., Hart, B.T., 2007. Examination of conflicts and improved strategies for the management of an endangered eucalypt species using Bayesian networks. Ecol. Model. 201, 37–59.

Pradhanang, S.M., Briggs, R.D., 2014. Effects of critical source area on sediment yield and streamflow. Water Environ. J. 28, 222–232.

Roos, T., Wettig, H., Grünwald, P., Myllymäki, P., Tirri, H., 2005. On discriminative Bayesian network classifiers and logistic regression. Mach. Learn. 59, 267–296.

Ropero, R.F., Aguilera, P.A., Rumí, R., 2015. Analysis of the socioecological structure and dynamics of the territory using a hybrid Bayesian network classifier. Ecol. Model. 311, 73–87.

Rumí, R., 2003. Modelos de redes Bayesianas con variables discretas y continuas. Ph.D. thesis. Universidad de Almer&rsquo;í.

Rumí, R., Salmerón, A., 2007. Approximate probability propagation with mixtures of truncated exponentials. Int. J. Approx. Reason. 45, 191–210.

Rumí, R., Salmerón, A., Moral, S., 2006. Estimating mixtures of truncated exponentials in hybrid Bayesian networks. Test 15, 397–421.

Schmitz, M., Pineda, F., Castro, H., Aranzabal, I.D., Aguilera, P., 2005. Cultural landscape and socioeconomic structure. In: Environmental value and demand for tourism in a Mediterranean territory. Consejer&rsquo;íde Medio Ambiente. Junta de Andaluc&rsquo;í, Sevilla.

Scott, M.W., 2010. Logistic Regression. From Introductory to Advanced Concepts and Applications. Sage.

Segurado, P., Araújo, M.B., 2004. An evaluation of methods for modelling species distribution. J. Biogeogr. 31, 1555–1568.

Shenoy, P.P., West, J.C., 2011. Inference in hybrid Bayesian networks using mixtures of polynomials. Int. J. Approx. Reason. 52 (5), 641–657.

Uusitalo, L., 2007. Advantages and challenges of Bayesian networks in environmental modelling. Ecol. Model. 203, 312–318.

van der Gaag, L.C., Renooij, S., Feelders, A., de Groote, A., Eijkemans, M.J.C., Broekmans, F.J., Fauser, B.C.J.M., 2009a. Aligning Bayesian network classifiers with medical contexts. In: Perner, P. (Ed.), Proceedings of the Sixth International Conference on Machine Learning and Data Mining in Pattern Recognition, vol. 5632 of Lecture Notes in Artificial Intelligence. Springer-Verlag, Berlin, Heidelberg, pp. 787–801.

van der Gaag, L.C., Renooij, S., Steeneveld, W., Hogeveen, H., 2009b. When in doubt . . . be indecisive. In: Sossai, C., Chemello, G. (Eds.), Proceedings of the European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty, vol. 5590 of Lecture Notes in Artificial Intelligence. Springer-Verlag, Berlin, Heidelberg, pp. 518–529.

Voinov, A., Bousquet, F., 2010. Modelling with stakeholders. Environ. Model. Softw. 24, 1268–1281.

Yang, Y., Webb, G., 2009. Discretization for naive-Bayes learning: managing discretization bias and variance. Mach. Learn. 74, 39–74.

Yang, Y., Webb, G., Wu, X., 2010. Discretization methods. In: Data Mining and Knowledge Discovery Handbook. Springer, pp. 101–116.

Zhou, Y., Fenton, N., Neil, M., 2014. Bayesian network approach to multinomial parameter learning using data and expert judgments. Int. J. Approx. Reason. 55, 1252–1268.