A Bayesian naïve Bayes classifier for dating archaeological sites

Carmen Armero[1], Gonzalo García-Donato[2], Joaquín Jiménez-Puerto[1], Salvador Pardo-Gordó[1], Joan Bernabeu[1]

[1] Universitat de València, Spain
[2] Universidad de Castilla-La Mancha, Spain

E-mail for correspondence: carmen.armero@uv.es

Abstract: Dating is a key element for archaelogists. We propose a Bayesian approach to provide chronology to sites that have neither radiocarbon dating nor clear stratigraphy and whose only information comes from bifacial flint arrowheads. This classifier is based on the Dirichlet-multinomial inferential process and posterior predictive distributions. The procedure is applied to predict the period of a set of undated sites located in the east of the Iberian Peninsula during the IVth and IIIrd millennium cal. BC

Keywords: Bifacial flint arrowheads; Dirichlet-multinomial process; Posterior predictive distribution

1 Introduction

Dating is a key element for archaeologists because they need a time scale to locate the information collected from the excavations and field work in order to build, albeit with uncertainty, our most remote past. Archaeological scientists generally use stratigraphic expert information and dating techniques for examining the age of the relevant artifacts. Bayesian inference is commonly used in archaeology as a tool to construct robust chronological models based on information from scientific data as well as expert knowledge (e.g. stratigraphy) (Buck el at., 1996).
Radiocarbon dating is one of the most popular techniques for obtaining data due to its presence in any being that has lived on Earth. However, it is not always possible in all studies to collect organic material and obtain that type of data or to have good stratigraphic references. In this context, we propose a Bayesian approach to provide chronology to some archaeological

This paper was published as a part of the proceedings of the 35th International Workshop on Statistical Modelling (IWSM), Bilbao, Spain, 19–24 July 2020.

sites that do not have radiocarbon dates and show unprecise stratigraphic relationships.
We propose an automatic Bayesian procedure, very popular in text classification (Wang *et al.*, 2003), based on predictive probability distributions, for classifying the period to which an undated site belongs based on the type and number of arrows found in it. This proposal takes into account on the Dirichlet-multinomial inferential process for learning about the proportion of different types of arrowheads in each chronological period and the concept of posterior predictive distribution for a new undated site. This procedure is applied to date a set of sites located in the east of the Iberian Peninsula during the IVth and IIIrd millennium cal. BC. During this time, bifacial flint arrowheads appear and spread. Archaeological research suggests that the shape of these arrowheads could be related with specific period and/or geographical social units spatially defined.

2 Bayes classifier

The prediction of the period to which an undated site belongs based on information about the number and type of arrows that have been collected at this site includes two different phases.

2.1 Dirichlet-multinomial inferential process

Let Y_{ij} be the random variable that describes the number of type j arrowheads, of the total n_i collected, in the sites belonging to period i, $i = 1, \ldots, I$, and consider $\boldsymbol{Y}_i = (Y_{i1}, Y_{i2}, \ldots, Y_{iJ})'$. A probabilistic model for $\boldsymbol{Y}_i \mid \boldsymbol{\theta}_i$ is the multinomial distribution, $\text{Mn}(\boldsymbol{\theta}_i, n_i)$, where $\boldsymbol{\theta}_i = (\theta_{i1}, \theta_{i2}, \ldots, \theta_{iJ})'$ is a probability vector and θ_{ij} is the probability that an arrowheads of period i is of type j.
We assume a Perks' prior distribution (Armero et *al.*, 2018) for $\boldsymbol{\theta}_i$. The subsequent posterior posterior distribution is the Dirichlet (Dir) distribution

$$\pi(\boldsymbol{\theta}_i \mid \mathcal{D}_i) = \text{Dir}(\alpha_{i1} = y_{i1} + (1/J), \ldots, \alpha_{iJ} = y_{iJ} + (1/J))$$

where y_{ij} is number of arrowheads of type j in the period i and $\mathcal{D}_i = \{y_{i1}, \ldots, y_{iJ}\}$. The marginal posterior distribution for each probability θ_{ij} ia beta distribution $\text{Be}(\alpha_{ij}, \alpha_{i+} - \alpha_{ij})$, with $\alpha_{i+} = \sum_{j=1}^{J} \alpha_{ij}$.

2.2 Classification process

After learning about the distribution of the number of arrowheads types in each site, we have to assign a period m^* to a new site with a given number and type of arrowsheads recorded. We consider a new undated

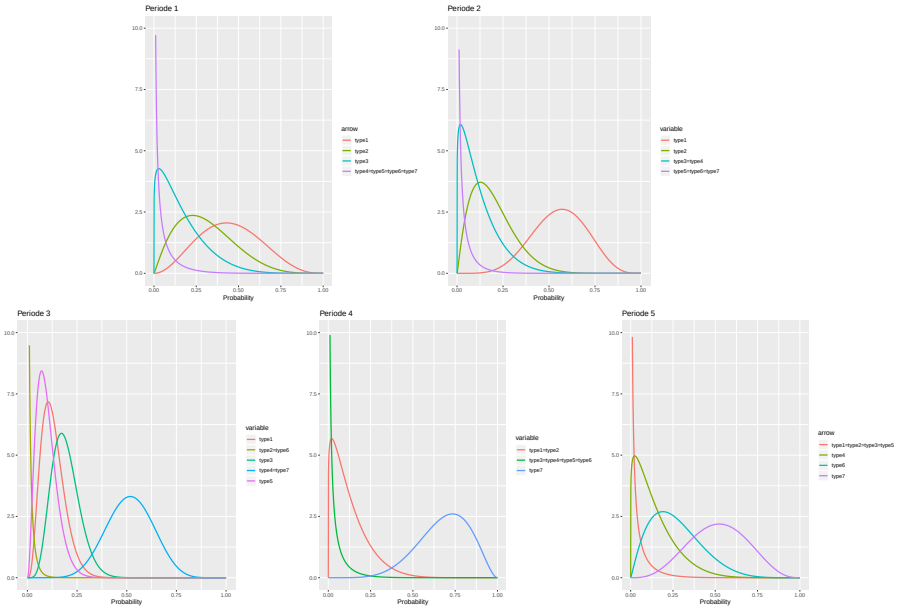
site s^* in which we found a total of n^* arrowheads distributed by type according to $\boldsymbol{y}^* = (y_1^*, \ldots, y_J^*)$. The relevant scientific question is now about the probability that this site belongs to each of the different time periods considered. Following Bayes' theorem:

$$P(m^* = m_i \mid \boldsymbol{y}^*, \mathcal{D}) \propto P(\boldsymbol{y}^* \mid m^* = m_i, \mathcal{D})\, P(m^* = m_i \mid \mathcal{D})$$

where $\mathcal{D} = \cup \mathcal{D}_i$, $(\boldsymbol{y}^* \mid m^* = m_i, \mathcal{D})$ follows a Dirichlet-multinomial distribution $\text{DiMn}(n^*, \boldsymbol{\alpha}_i)$ with $n^* = \sum y_j^*$, and $P(m^* = m_i \mid \mathcal{D})$ can be estimated as the proportion of sites in the sample for each of the periods under consideration.

3 East of the Iberian Peninsula sites during the IVth and IIIrd millennium cal. BC.

Five chronological periods in the east of the Iberian Peninsula sites during the IVth and IIIrd millennium cal. BC. were studied. They include arrowheads data from several archaeological contexts, *Niuet*, *Jovades 1* and *Jovades 2* from period 1, *Quintaret*, *Jovades 3*, *Jovades 4*, and *Niuet 2* from period 2, *Migdia 1*, *Beniteixir* , *La Vital 1* , *Randero 1*, *Niuet 3*, *Niuet 4*, and *Diablets* from period 3, *Migdia 2*, *Missena 1*, and *La Vital 2* from period 4, and *Arenal costa*, *Missena 2*, and *La Vital 3* from period 5.



The Figure above shows the posterior marginal distribution of the abundance of the different types of arrowheads in each of the five chronological

periods considered. Type 1 and 2 arrowheads are most abundant in periods 1 and 2, with an increase in type 1 compared to type 2 arrowheads in the second period. During period 3, type 4 and type 7 arrowheads are more abundant. The latter are clearly the most used in period 4, which become less used in period 5 when type 6 arrowheads appears with more probability.
The posterior probability that a new site belongs to each of the periods considered was estimated as 0.15 for periods 1, 4 and 5, 0.20 for period 2, and 0.35 for period 3.
The following table presents the posterior predictive distribution of the period to which a series of new undated sites belong, whose only available information is based on the number and type of arrows found collected.

Site	Period 1	Period 2	Period 3	Period 4	Period 5
Rambla C.	0.0001	0.0000	0.0004	0.9339	0.0654
Ereta I	0.7804	0.2196	0.0000	0.0000	0.0000
Ereta II	0.5019	0.4901	0.0000	0.0075	0.0005
Ereta III	0.0694	0.0912	0.8330	0.0060	0.0004
Ereta IV	0.0021	0.0098	0.6358	0.3504	0.0019

The results obtained present a great agreement with the expert information of the archaeologists of the project, so it is a proposal that can be very useful in archaeological research.

References

Alvares, D., Armero, C., and Forte, A. (2018). What Does Objective Mean in a Dirichlet-multinomial Process? *International Statistical Review*, **86**, 106 – 118.

Buck, I. C. E., Cavanagh, W. G., and Litton, C. D. (1996). *Bayesian Approach to Interpreting Archaeological Data.* Chiscester: Wiley.

Wang, Y., Hodges, J. and Tang, B. (2003). Classification of Web Documents Using a Naive Bayes Method. *15th IEEE International Conference on Tools with Artificial Intelligence*, **124**, 560 – 564.