International Journal of Approximate Reasoning 55 (2014) 1548–1569

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

www.elsevier.com/locate/ijar

# Imprecise probability models for learning multinomial distributions from data. Applications to learning credal networks

Andrés R. Masegosa, Serafín Moral *

*Dpto. Ciencias de la Computación e Inteligencia Artificial, Universidad de Granada, 18071 Granada, Spain*

ARTICLE INFO

*Article history:*
Available online 1 October 2013

*Keywords:*
Imprecise probability
Learning
Ignorance
Imprecise prior models
Credal networks

ABSTRACT

This paper considers the problem of learning multinomial distributions from a sample of independent observations. The Bayesian approach usually assumes a prior Dirichlet distribution about the probabilities of the different possible values. However, there is no consensus on the parameters of this Dirichlet distribution. Here, it will be shown that this is not a simple problem, providing examples in which different selection criteria are reasonable. To solve it the Imprecise Dirichlet Model (IDM) was introduced. But this model has important drawbacks, as the problems associated to learning from indirect observations. As an alternative approach, the Imprecise Sample Size Dirichlet Model (ISSDM) is introduced and its properties are studied. The prior distribution over the parameters of a multinomial distribution is the basis to learn Bayesian networks using Bayesian scores. Here, we will show that the ISSDM can be used to learn imprecise Bayesian networks, also called credal networks when all the distributions share a common graphical structure. Some experiments are reported on the use of the ISSDM to learn the structure of a graphical model and to build supervised classifiers.



## 1. Introduction

The basic problem we consider in this paper is the estimation of the parameters of a multinomial distribution from data. Assume a variable $X$ taking values on a finite set $W = \{w_1, \ldots, w_k\}$ for which we have a set of independent and identically distributed observations $\mathcal{D} = (D_1, \ldots, D_n)$. We want to estimate the probability distribution of $X$. We consider the parameter set:

$$\Theta = \left\{ \theta = (\theta_1, \ldots, \theta_k) \;\middle|\; \theta_j \geqslant 0,\ \forall j,\ \sum_{j=1}^{k} \theta_j = 1 \right\},$$

where $P(X = w_j|\theta) = \theta_j$. The value $\theta_j$ is called the *chance* associated with $w_j$. If $P$ is a probability measure about $X$, then it associated probability distribution will be denoted by $p$. In the paper, a probability distributions $p(w_i|\theta)$ will be sometimes denoted as $p(X = w_j|\theta)$ to make explicit the corresponding variable.

We assume that we do not have any other information about the parameter set, and we want to make inferences about $\theta$ based only on the data $\mathcal{D} = (D_1, \ldots, D_n)$ as in Ref. [1].

* Corresponding author.
*E-mail addresses:* andrew@decsai.ugr.es (A.R. Masegosa), smc@decsai.ugr.es (S. Moral).

0888-613X/$ – see front matter 
http://dx.doi.org/10.1016/j.ijar.2013.09.019

The objective Bayesian approach [2,3] tries to assess a prior probability density in the parameter space $\Theta$ representing the state of ignorance. The idea is to determine it by assuming some rationality principles that should be satisfied [4], as the *symmetry principle* saying that, under ignorance, the density should be invariant by permutations of the parameters. Unfortunately, there is not an agreement about which prior probability is the most reasonable; also, sometimes so many rationality principles are imposed that there is not a single probability satisfying all of them [5]. The most usual prior density is the Dirichlet distribution in $\Theta$ but there is not a unique criterion for selecting the *equivalent sample size* parameter.

In this situation, it seems natural to use imprecise probability models [6] and to assume a set of possible prior distributions instead of a single one, in contrast with the pure Bayesian methodology that assumes a single prior distribution. The first serious attempts to build a theory of imprecise probability go back to Keynes [7] but it is after the publication of Walley's book [6] when the theory is experimenting a more intense and unified development. The basic idea is that to represent a lack of knowledge about the parameters an imprecise representation is much more appropriate than a precise one.

An example of this approach is the Imprecise Dirichlet Model (IDM) by Walley [8,9]. Inferences are obtained by assuming a family of Dirichlet distributions for them with a fixed equivalent sample size. This model has some nice properties (e.g., it verifies the *representation invariance principle* [8,10]) and has been applied in a variety of different situations [11–15]. But in some cases it produces too uninformative conclusions, for example when the observations are imperfect (they can be erroneous) and they only induce a likelihood on $\Theta$ [16,17]. Furthermore, it is not useful to decide about independence relationships between variables with generalized Bayesian scores [18]. This makes impossible to apply the general IDM to learn graphical models representing independence relationships between variables. Another model recently proposed for learning from a sequence of observations is based on the *non-parametric predictive inference* methodology [19]. However, this model is not based on a set of prior distributions about the parameters which is updated by conditioning to the observations, but on a post-data assumption about the uncertainty associated with a future observation.

In this paper we study an alternative imprecise model to the IDM. It is based on sets of symmetrical prior Dirichlet distributions, but with different equivalent sample sizes. This idea can be found in Ref. [6, Section 5.4] as a model in which imprecision arises from the *degree of conflict* between the set of prior distributions and the frequencies found in the observed data. It will be called the *Imprecise Sample Size Dirichlet Model* (ISSDM). Some of the properties of the IDM will be lost (representation invariance and the initial vacuous intervals under no observations), but it will be more powerful in those cases where the inferences produced by the IDM seems to be too weak. To make this fact more relevant, we introduce the *learning principle* as a requirement for any imprecise prior model and which is satisfied by the ISSDM but not by the IDM.

The importance of the selection of the equivalent sample size is especially relevant when learning Bayesian networks [20]. The use of Bayesian scores to learn the structure of a Bayesian network and the methods for estimating the parameters are also based on assuming a prior Dirichlet distribution for the values of the conditional probabilities of each node given its parents. But there is an additional problem when learning conditional probabilities: the equivalent sample size should also change with the number of conditional probabilities we are estimating for one variable. The nature of this change will be a new source for imprecision in the determination of the equivalent sample size. In this paper we will try to show that the ISSDM can be used to overcome some of the difficulties associated with Bayesian learning, being the result a generalized *credal network* [21], a graphical model in which the structure and the conditional probabilities can be imprecise.

This paper is organized in the following way. In Section 2 we describe the Bayesian approach and its limitations. In Section 3 the IDM is presented and we give examples showing weak behavior of this model related with conditioning. In Section 4 we introduce the learning principle, the ISSDM and discuss its properties. Section 5 shows how the ISSDM can be applied to learn generalized credal networks, while Section 6 will show some experiments. Finally, the conclusions will be given in Section 7.

## 2. The Bayesian approach

The Bayesian approach usually assumes a prior Dirichlet distribution $D(\alpha_1, \ldots, \alpha_k)$ on $\Theta$ and makes inferences by conditioning the prior distribution to the data. The Dirichlet density is as follows:

$$f(\theta) = \frac{\prod_{i=1}^{k} \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{k} \alpha_i)} \prod_{i=1}^{k} \theta_i^{\alpha_i - 1},$$

where the parameters $\alpha_i$ are all positive, and $\Gamma$ is the gamma function. The sum $\sum_{i=1}^{k} \alpha_i$ is called the *equivalent sample size* and is denoted by $s$. We will call $\alpha_i$ the *prior weight* for value $w_i$.

One of the reasons to assume a Dirichlet distribution is its computational simplicity as it is a conjugate density of the multinomial distribution. This means that the posterior density is also Dirichlet, specifically it is a Dirichlet $D(\alpha_1 + n_1, \ldots, \alpha_k + n_k)$ where $n_i$ is the number of times that $[X = w_i]$ has been observed in data $\mathcal{D}$. Observe that the equivalent sample size has increased by $n = \sum_{i=1}^{k} n_i$ with respect to the prior sample size, i.e. after conditioning to a sample of size $n$ the equivalent sample size increases by $n$.

The estimation of $\theta$ under quadratic loss is the expected value of the posterior distribution, which is easily computed as:

$$p(X = w_i|\mathcal{D}) = \hat{\theta}_i = P(\theta_i|\mathcal{D}) = \frac{n_i + \alpha_i}{n + s},$$