European Journal of Operational Research 212 (2011) 112–122

Contents lists available at ScienceDirect

European Journal of Operational Research

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

Stochastics and Statistics

Maximising entropy on the nonparametric predictive inference model for multinomial data

Joaquín Abellán [a,*], Rebecca M. Baker [b], Frank P.A. Coolen [b]

[a] Department of Computer Science and Artificial Intelligence, University of Granada, Granada, Spain
[b] Department of Mathematical Sciences, Durham University, Durham, UK

ARTICLE INFO

Article history:
Received 7 April 2010
Accepted 10 January 2011
Available online 22 January 2011

Keywords:
Entropy maximization
Imprecise probabilities
Nonparametric predictive inference
Uncertainty measures

ABSTRACT

The combination of mathematical models and uncertainty measures can be applied in the area of data mining for diverse objectives with as final aim to support decision making. The maximum entropy function is an excellent measure of uncertainty when the information is represented by a mathematical model based on imprecise probabilities. In this paper, we present algorithms to obtain the maximum entropy value when the information available is represented by a new model based on imprecise probabilities: the nonparametric predictive inference model for multinomial data (NPI-M), which represents a type of entropy-linear program. To reduce the complexity of the model, we prove that the NPI-M lower and upper probabilities for any general event can be expressed as a combination of the lower and upper probabilities for the singleton events, and that this model can not be associated with a closed polyhedral set of probabilities. An algorithm to obtain the maximum entropy probability distribution on the set associated with NPI-M is presented. We also consider a model which uses the closed and convex set of probability distributions generated by the NPI-M singleton probabilities, a closed polyhedral set. We call this model A-NPI-M. A-NPI-M can be seen as an approximation of NPI-M, this approximation being simpler to use because it is not necessary to consider the set of constraints associated with the exact model.

1. Introduction

When we have a set of data about individual items, including characteristics and a variable in study, it is important the model (normally a mathematical one) used to represent the information that the set of data give us. To quantify the information that a characteristic, or a set of them, represents about the variable in study, we can use measures of information-based uncertainty (for simplicity we call these *uncertainty measures*) on the mathematical model used.

Mathematical models and uncertainty measures can be combined to present procedures into the areas of supervised classification learning, variable selection methods, clustering, etc. All of them are into the general area of data mining, and can be considered as important tools for decision support. References about this type of procedures are [2,4,5,8].

There are many mathematical models that can be used to represent the information without single-valued probabilities. These models are generalisations of probability theories such as belief functions, reachable probability intervals, capacities of various orders, upper and lower probabilities, and closed and convex sets of probability distributions (also called credal sets). The term *imprecise probability* (Klir [30], Walley [39]) subsumes these theories. Some of these generalised theories are more appropriate than others in specific situations.

The Imprecise Dirichlet model (IDM), presented by Walley [40], is a mathematical model for statistical inference from multinomial data which was developed to correct shortcomings of previous objective models. It verifies a set of principles which are claimed by Walley to be desirable for inference (see Walley [40]). The IDM can be seen as a model which gives imprecise probabilities that can be expressed via a set of reachable probability intervals and a belief function (Abellán [1]). The IDM has been applied to various statistical problems and a description of these applications was presented by Bernard [12]. However, the use of the IDM has recently been questioned for some practical applications (Piatti et al. [34]). Shortcomings of the IDM were already discussed in detail by Walley [40], and by many discussants of that paper, leading Walley to strongly motivate researchers to develop alternative models for such inference.

Coolen and Augustin [10,17] presented nonparametric predictive inference for multinomial data (NPI-M) as such an alternative, which does not suffer from some of the main drawbacks of the IDM [40]. It is different from the IDM in the sense that NPI-M learns from data in the absence of prior knowledge and with relatively few modelling assumptions, most noticeably a post-data

* Corresponding author. Tel.: +34 958242376.
E-mail addresses: jabellan@decsai.ugr.es (J. Abellán), r.m.baker@durham.ac.uk (R.M. Baker), frank.coolen@durham.ac.uk (F.P.A. Coolen).

0377-2217/$ - see front matter
doi:10.1016/j.ejor.2011.01.020

exchangeability-like assumption together with a latent variable representation of data as lines on a probability wheel. NPI-M does not satisfy all of the principles for inference suggested by Walley [40], specifically the Representation Invariance Principle (RIP), but Coolen and Augustin do not consider this to be a shortcoming [18]. In fact, they present strong arguments against general adoption of the RIP for inference and propose an alternative, weaker principle, which NPI-M satisfies.

With the emergence of models that extend classical probability theory, an extension of information-based uncertainty theory has been needed. In the 1990s, using the Shannon entropy (Shannon [37]) measure for probabilities as a starting point, a large amount of research was carried out to study measures to quantify different types of uncertainty inherent in belief functions. In recent years, this study has been extended to general credal sets. The maximum entropy measure appears as a suitable total uncertainty measure for general credal sets, verifying a set of desirable properties (Abellán et al. [6], Klir [30]). This measure have been questioned as an aggregate measure into some theories ([3,29,31]), where the efficiency of its calculus is very important ([33,38]). Since Jaynes [26–28] expounded his principle of maximum entropy, this measure has been extensively used in the literature.

In this paper, we study the NPI-M model with a view to practical applications, principally in data mining. With this aim in mind, we consider the applications in [8,2,4], where the IDM is applied with the maximum entropy measure. NPI-M is an alternative model for uncertainty quantification that can replace the IDM in some situations where the use of the IDM is questioned. The inferences given by NPI-M are in the form of lower and upper probabilities for events. We prove that these bounds comprise sets of reachable probability intervals, which enables more efficient computation (Campos et al. [13]).

An important characteristic of the set of probability distributions generated by NPI-M is that it is not a closed and convex set. We can determine bounds on the probability of each general event via the probabilities of the singleton events, but not all of the distributions in the associated credal set are compatible with the theoretical NPI-M model. For ease of application, an approximation of this model can be used, specifically the use of the credal set associated with the set of reachable probability intervals rather than the actual set of distributions valid under NPI-M. This approximation will be referred to as A-NPI-M.

When working with reachable probability intervals, several algorithms can be used to find the maximum entropy algorithm within the associated credal set, such as the algorithm presented by Abellán and Moral [7] for probability intervals or the more general algorithm presented by Abellán and Moral [9]. However, these algorithms can not be used with NPI-M because due to the set of constraints of the model it does not give a closed and convex set of probability distributions. Taking into account all constraints of NPI-M, we present an algorithm to obtain the maximum entropy distribution on the set of probability distributions generated by NPI-M.

Finally, we present an efficient algorithm to obtain the maximum entropy distribution on the set generated by A-NPI-M. This is simpler than the NPI-M algorithm and we can base it on the algorithm presented by Abellán and Moral [7] because A-NPI-M generates a closed and convex set of probability distributions.

This paper is organised as follows: in Section 2, we present a summary of the principal theories of imprecise probability; in Section 3, we explain NPI-M; Section 4 is devoted to an overview of uncertainty measures; in Section 5 we present an algorithm to calculate the maximum entropy probability distribution on the set obtained from NPI-M; in Section 6, we present an algorithm to calculate the maximum entropy probability distribution on the set obtained from A-NPI-M; and finally, Section 7 is devoted to our conclusions.

2. Imprecise probability theories: a brief overview

2.1. Credal sets

Theories of imprecise probability (Klir [30], Walley [39], Weichselberger [43]) share some common characteristics: for example, the evidence within each theory can be described by a lower probability function P_* on a finite set[1] X or, alternatively, by an upper probability function P^* on X. These functions are always regular monotone measures (Wang and Klir [41]) and satisfy

$$\sum_{x\in X} P_*(\{x\}) \leqslant 1, \quad \sum_{x\in X} P^*(\{x\}) \geqslant 1. \tag{1}$$

If the set of probability distributions comprises a general credal set, $\mathcal{P}$, i.e. a closed and convex set of probability distribution functions p, on a finite set X (Kyburg [32]), functions P_* and P^* associated with $\mathcal{P}$ are determined for each set $A \subseteq X$ by the expressions

$$P_*(A) = \inf_{p\in\mathcal{P}} \sum_{x\in A} p(\{x\}), \quad P^*(A) = \sup_{p\in\mathcal{P}} \sum_{x\in A} p(\{x\}). \tag{2}$$

P_* and P^* are called dual, because for each $p \in \mathcal{P}$ and each $A \subseteq X$, the following holds:

$$P^*(A) = 1 - P_*(X - A), \tag{3}$$

where $X - A$ denotes the subset of X that is complementary to A.

Any given lower probability function P_* is uniquely represented by a set-valued function m for which $m(\emptyset) = 0$ and

$$\sum_{A\in\wp(X)} m(A) = 1, \tag{4}$$

where $\wp(X)$ is the power set of X (Chateauneuf and Jaffray [14]; Grabisch [20]).

Any set $A \subseteq X$ for which $m(A) \neq 0$ is often called a focal element, and the set of all focal elements with the values assigned to them by function m is called a body of evidence. The function m is called a Möbius representation of P_* when it is obtained for all $A \subseteq X$ via the Möbius transform

$$m(A) = \sum_{B|B\subseteq A} (-1)^{|A-B|} P_*(B). \tag{5}$$

The inverse transform is defined for all $A \subseteq X$ by the formula

$$P_*(A) = \sum_{B|B\subseteq A} m(B). \tag{6}$$

It follows directly from (5) that

$$P^*(A) = \sum_{B|B\cap A\neq\emptyset} m(B), \tag{7}$$

for all $A \subseteq X$. Assume now that evidence is expressed in terms of a given lower probability function P_*. Then the set of probability distribution functions $\mathcal{P}(P_*)$ that is consistent with P_*, which is always a closed and convex set, is defined as follows:

$$\mathcal{P}(P_*) = \left\{ p | x\in X; p(x)\in[0,1]; \sum_{x\in X} p(x) = 1; P_*(A) \leqslant \sum_{x\in A} p(x); \forall A\subseteq X \right\}. \tag{8}$$

2.2. Choquet capacities of various orders

An important theory of imprecise probability is the theory of Choquet capacities of various orders (Choquet [15]). The most general theory in this category is the theory based on capacities of

[1] Or a finite variable X, with values in a finite set $Z = \{x_1, \ldots, x_K\}$.