Computational Statistics and Data Analysis 71 (2014) 789–802

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Classification with decision trees from a nonparametric predictive inference perspective

Joaquín Abellán [a,*], Rebecca M. Baker [b], Frank P.A. Coolen [b], Richard J. Crossman [c], Andrés R. Masegosa [a]

[a] *Department of Computer Science and Artificial Intelligence, University of Granada, Granada, Spain*

[b] *Department of Mathematical Sciences, Durham University, Durham, UK*

[c] *Warwick Medical School, University of Warwick, Coventry, UK*

ARTICLE INFO

Article history:
Received 8 September 2011
Received in revised form 7 February 2013
Accepted 7 February 2013
Available online 27 February 2013

Keywords:
Imprecise probabilities
Imprecise Dirichlet model
Nonparametric predictive inference model
Uncertainty measures
Supervised classification
Decision trees

ABSTRACT

An application of nonparametric predictive inference for multinomial data (NPI) to classification tasks is presented. This model is applied to an established procedure for building classification trees using imprecise probabilities and uncertainty measures, thus far used only with the imprecise Dirichlet model (IDM), that is defined through the use of a parameter expressing previous knowledge. The accuracy of that procedure of classification has a significant dependence on the value of the parameter used when the IDM is applied. A detailed study involving 40 data sets shows that the procedure using the NPI model (which has no parameter dependence) obtains a better trade-off between accuracy and size of tree than does the procedure when the IDM is used, whatever the choice of parameter. In a bias-variance study of the errors, it is proved that the procedure with the NPI model has a lower variance than the one with the IDM, implying a lower level of over-fitting.

1. Introduction

Using mathematical models to represent the information available is a common practice, including for those situations where uncertainty is of concern. Many models are designed for such situations, and generalise probability theories such as belief functions, reachable probability intervals, capacities of various orders, upper and lower probabilities, and closed convex sets of probability distributions (also called *credal sets*). These generalisations, some of which are more appropriate than others in specific situations, are subsumed within the term *imprecise probability* (Klir, 2006; Walley, 1991).

Developing such extensions of probability theory meant classical uncertainty-based information theory needed likewise extending. In the 1990s, using the Shannon's entropy measure (Shannon, 1948) for probabilities as a starting point, a large amount of research was carried out to present measures which would quantify different types of uncertainty inherent to some of these models. At first this was done principally with respect to belief functions, but in recent years the study has been extended to general credal sets. The maximum entropy measure has been used as a suitable total uncertainty measure for general credal sets, satisfying a number of desirable properties (Abellán et al., 2006; Klir, 2006). This paper focuses on applying imprecise probability models to the field of classification, through the maximisation of entropy over sets of probabilities.

One such model is the *imprecise Dirichlet model* (Walley, 1996), which makes statistical inferences from multinomial data. It was developed to overcome shortcomings in previous alternative objective models, such as for example the difficulty in

* Corresponding author. Tel.: +34 958242376; fax: +34 958240079.
E-mail addresses: jabellan@decsai.ugr.es (J. Abellán), r.m.baker@dunelm.org.uk (R.M. Baker), frank.coolen@durham.ac.uk (F.P.A. Coolen), r.j.crossman@warwick.ac.uk (R.J. Crossman), andrew@decsai.ugr.es (A.R. Masegosa).

0167-9473/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.csda.2013.02.009

justifying any particular prior in circumstances of total ignorance. The IDM satisfies a set of principles which are claimed (Walley, 1996) to be desirable for inference, most particularly the *representation invariance principle* (RIP) which states that inferences on future events should be independent of the arrangement and labelling of the sample space. The RIP states that an inference with multinomial data which involves the event that an observation belongs to a certain category, should not depend on how the other categories are represented. This implies, for example, that a binary situation, with all other categories combined together into a single category, should not lead to different inferences than with all other categories represented individually.

The IDM has been applied to many statistical problems; a description of these applications can be found in Bernard (2005). However, the appropriateness of the IDM for some practical applications has recently been questioned Piatti et al. (2005) on the grounds that the method does not deal well with incorrectly identified observations. Further shortcomings of the IDM in other situations were already discussed in detail by Walley (1996), and by many discussants of that paper, leading Walley to strongly motivate researchers to develop alternative inference models.

Coolen and Augustin (2005) presented nonparametric predictive inference for multinomial data (NPI-M) as an alternative approach that does not include some properties of the IDM which can cause problems in certain situations (Walley, 1996), such as (for example) the suggestion that with complete prior ignorance the second observation will be identical to the first with a probability that is frequently at least $\frac{1}{3}$, or that the probability of observing a new category is independent of the number of categories already seen. The NPI-M learns from data in the absence of prior knowledge and with only a few modelling assumptions, most noticeably a post-data exchangeability-like assumption together with a latent variable representation of data as lines on a probability wheel. NPI-M does not satisfy some of the principles for inference suggested by Walley (1996), most noticeably the RIP, instead satisfying an alternative, weaker principle. Several reasons are brought forward which put doubt about the general acceptance of the RIP for inference with imprecise probabilities. Most importantly perhaps is the result that any inferences on as yet unobserved categories would not depend on how many categories have been observed. Coolen and Augustin (2005) argue strongly in favour of a weaker assumption, allowing imprecision to increase if the data representation for multinomial observations exists of more categories, in which case a binary representation leads to minimal imprecision. The NPI approach follows this weaker assumption.

In addition to the data, the IDM is determined by a single parameter s; most commonly $s \in [1, 2]$. The larger the value of s, the less the probability intervals are altered by new observations. The most frequent value for the s parameter for the IDM model is $s = 1$, and when this value is used, the resulting imprecision is smaller than that expressed by the NPI model (NPI-M). Hence, it is important to consider the differences in the total uncertainty expressed by the maximum entropy function between the IDM and the NPI-M.

An application of the study of information based uncertainty measures with imprecise probabilities is the method of Abellán and Moral (2003a) for building decision trees. This method uses a split criterion with different characteristics to those of the classical split criteria, and is able to use both different imprecise probability models and different uncertainty measures when constructing a tree representing the information from a data set. The application of the IDM and the NPI-M to this approach produces in general different trees because, as we will see, the treatment of information is different in each case, with greater imprecision when the NPI-M is used.

It must be remarked that the accuracy of the above mentioned procedure with the IDM has a high dependence of the s parameter used. For some data sets, it is possible to obtain better results in accuracy when lower values of s are used (close to 0); and for others data sets larger values of the parameter (upper 2) give us better results. Thus far, we do not know the relation among the characteristics of a data set and the value of the parameter used.

We have carried out a series of experiments in order to check the performance of this approach when it uses the NPI-M to build decision trees. To this end, algorithms to attain the maximum entropy probability are required; these are presented in Abellán et al. (2011). We have used 40 data sets with the common characteristic that the class variable has a known number $K \geq 3$ of cases or categories, just as was considered in the model presented in Coolen and Augustin (2005). Our results are compared here with another classical split criterion and with the IDM with various values of s.

We will show that applying the NPI-M leads to slightly greater accuracy than the best model based on the IDM with $s = 1$, whilst generating notable smaller trees. By increasing the value of s in the IDM it is possible to generate similarly small trees, but in doing so the accuracy is noticeably decreased. Also, in a bias-variance decomposition of the errors we show that the procedure with the NPI-M presents a lower variance implying a lower level of over-fitting.

The paper is arranged as follows: Section 2 presents a summary of the principal theories of imprecise probabilities. Sections 3 and 4 describe the IDM and NPI-M, respectively. Section 5 gives a brief overview of uncertainty measures for imprecise probabilities. Section 6 explains the procedure for building decision trees using imprecise probabilities and uncertainty measures, and in Section 7 we present experimental results. Conclusions are given in Section 8.

2. Imprecise probabilities: a brief overview

2.1. Imprecise probabilities and credal sets

Most theories of imprecise probabilities (Klir, 2006; Walley, 1991; Weichselberger, 2001) share some common characteristics; for example, that the evidence within each theory can be described by a lower probability function P_* on a finite