Direct Nonparametric Predictive Inference Classification Trees

Abdulmajeed Alharbi ABDULMAJEED.A.ALHARBI@DURHAM.AC.UK
Frank P. A. Coolen FRANK.COOLEN@DURHAM.AC.UK
Tahani Coolen-Maturi TAHANI.MATURI@DURHAM.AC.UK
Department of Mathematical Sciences, Durham University, Durham, UK

Classification is one of the most common data mining techniques that is used for assigning a new observation to one of a set of predefined categories based on the attributes of the observation. There are many classification methods available in the literature, the classification tree is one of the most commonly used because of its interpretational simplicity. Several algorithms have been introduced in the literature to build classification trees. The C4.5 and the CART algorithms are two of the most commonly used classical algorithms. Classification trees are also built using imprecise probabilities, based on the imprecise Dirichlet model (IDM) [1] or Nonparametric Predictive Inference model for multinomial data (NPI-M) [2]. Coolen and Augustin [4] developed the NPI model for multinomial data (NPI-M) as an alternative approach to the IDM [5]. NPI is a statistical method which learns from data in the absence of prior knowledge and uses only few modelling assumptions, enabled by the use of lower and upper probabilities to quantify uncertainty [3].

In this work, we propose a new algorithm to build classification trees using imprecise probabilities and based on the NPI-M, which we call the Direct Nonparametric Predictive Inference classification tree algorithm for multinomial data (D-NPI-M). The D-NPI-M classification algorithm uses *Correct Indication* (CI) as a split criterion. The CI is a new split criterion that is completely based on the NPI-M lower and upper probabilities and does not use any additional concepts such as entropy. The CI is simply about the indication the attribute variables will give. After computing the lower and upper probabilities of CI for all attribute variables, we aim at maximum probability for both the lower and upper probabilities of CI. As a first application of the D-NPI-M classification trees, we start with an experimental analysis on six real-world datasets in order to examine the performance of the D-NPI-M algorithm when building classification trees. We also compare the performance of the D-NPI-M algorithm to classical classification tree algorithms such as the C4.5 and the CART algorithms, and imprecise classification algorithms based on the IDM or the NPI-M models. Initial results from this experimentation suggest that the D-NPI-M classification algorithm slightly outperforms other classification algorithms in terms of percentages of correct classifications.

References

[1] J. Abellán and S. Moral. Building classification trees using the total uncertainty criterion. *International Journal of Intelligent Systems*, 18(12):1215–1225, 2003.

[2] J. Abellán, R.M. Baker, F. P.A. Coolen, R. J. Crossman, and A.R. Masegosa. Classification with decision trees from a nonparametric predictive inference perspective. *Computational Statistics & Data Analysis*, 71:789–802, 2014.

[3] T. Augustin and F. P.A. Coolen. Nonparametric predictive inference and interval probability. *Journal of Statistical Planning and Inference*, 124(2):251–272, 2004.

[4] F. P.A. Coolen and T. Augustin. A nonparametric predictive alternative to the Imprecise Dirichlet Model: the case of a known number of categories. *International Journal of Approximate Reasoning*, 50(2):217–230, 2009.

[5] P. Walley. Inferences from multinomial data: learning about a bag of marbles. *Journal of the Royal Statistical Society: Series B*, 58(1):3–57, 1996.