Constructing Classification Trees with NPI-based Thresholds for Continuous Attributes

Masad Alrasheedi MASAD.A.ALRASHEEDI@DURHAM.AC.UK
Department of Mathematical Sciences, Durham University, UK
Department of Business Administration, Taibah University, SA

Tahani Coolen-Maturi TAHANI.MATURI@DURHAM.AC.UK
Frank Coolen FRANK.COOLEN@DURHAM.AC.UK
Department of Mathematical Sciences, Durham University, UK

In data mining, classification is a form of data analysis in which a machine learning algorithm assigns a specific category or class to new observations. Classification trees are considered one of the most widely used methods for classification tasks. They are constructed recursively from the top down using repeated splits of the training dataset. When this dataset includes continuous-valued attributes, binary splits are typically done by selecting the threshold value that maximises the splitting criterion used (e.g., C4.5 [5] gain ratio criterion or CART [2] Gini's index). In this work, we present the use of nonparametric predictive inference (NPI) for determining these thresholds of continuous attributes. NPI [1] is based on Hill's assumption $A_{(n)}$[4], which provides direct probabilities for future observations based on observed values of related random quantities.

The NPI-based thresholds method [3] has been developed for finding the best thresholds for two-groups and three-groups settings based on multiple future observations and the desired percentage value towards one group over another. For example, in a binary class, the values a and b be the desired proportions of correctly classified individuals in each group respectively. The process of choosing particular values of a and b depends on a person's beliefs of which class may be important to him to be correctly classified, for example, in healthcare, assume the values a and b is the desired percentage of correct classified from healthy and patients people respectively, one can take a value of b higher than a value, if giving the drug to patients is crucial, but this drug has no serious harmful effects on healthy people. It is expected that this will increase the proportion of the correct classification of diseased people more than healthy people. There is no constraint on these values, except to be in (0; 1]. So, the optimal threshold changes as changing the setting of these values which leads to a change in prediction accuracy. For this reason, setting meaningful target proportions for the predictive inference is considered.

In this work, we build the classification tree using the NPI-based thresholds for two-groups and three-groups classification. We propose using the optimisation technique (Genetic Algorithm (GA)) to discover the best setting of the desired percentage values based on the given data set. We use the lower probability for meeting these values criterion. For this reason, the work is presented at the conference on imprecise probabilities. Moreover, we use two levels of the ten-fold cross-validation procedure to verify the model performance. This model is applied to 10 real data sets and compared to classical methods such as C4.5 and CART. The results show that, in most cases, the proposed method leads to smaller classification trees with better accuracy compared to C4.5 and CART.

References

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

[2] Leo Breiman, Jerome Friedman, Charles Stone, and Richard Olshen. Classification and regression trees. *New York: Chapman & Hall*, 1984.

[3] Tahani Coolen-Maturi, Frank P.A. Coolen, and Manal Alabdulhadi. Nonparametric predictive inference for diagnostic test thresholds. *Communications in Statistics-Theory and Methods*, 49(3):697–725, 2020.

[4] Hill, Bruce M. Posterior distribution of percentiles: Bayes' theorem for sampling from a population. *Journal of the American Statistical Association*, 63(322):677–691, 1968.

[5] Ross Quinlan. C4.5: Program for machine learning. *Morgan Kaufmann Pub*, 1993.