A Bayesian approach to estimate probabilities in classification trees

Andrés Cano, Andrés R. Masegosa , Serafín Moral

Department of Computer Science and A.I.

University of Granada

1. Introduction

- Classification trees (CT) are one of the most used supervised classification models. But one of their main problems is the poor estimates of the class probabilities they produce [1].
- Good class probability estimates are essential in many tasks such as probability based ranking problems [2].
- This work proposes a Bayesian approach to build CT with excellent class probability estimates (CPE).

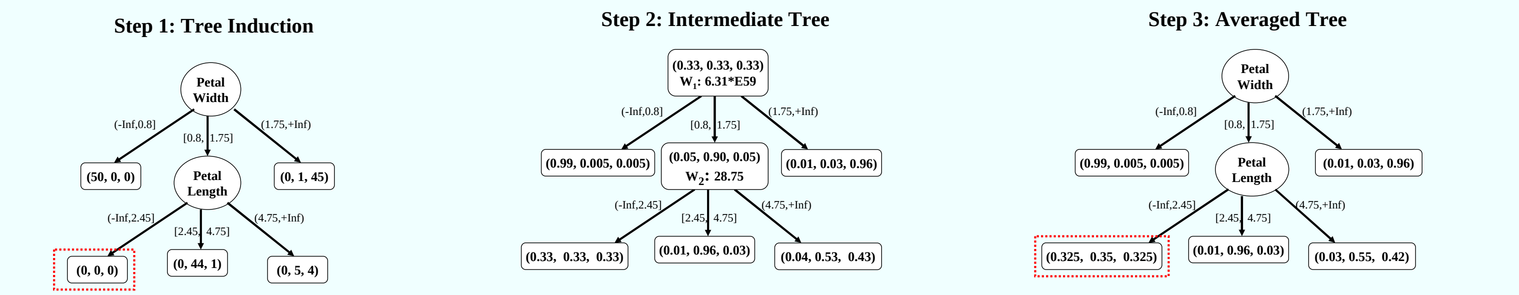
2. Bayesian Tree Induction (BTI)

- In this work, CT induction is faced as a Bayesian model selection problem [3].
- At each step it is selected the tree with MAP probability given the data. These options are evaluated:
  - Branch by a non-used node X in this branch: $P(M_X^t|\mathcal{D})$.
  - Stop the branching: $P(M^t|\mathcal{D})$.
- Eq. for selecting the splitting node or stop branching:

$$\frac{P(M_X^t|\mathcal{D})}{P(M^t|\mathcal{D})} = \frac{\prod_j^{|X|} \frac{\Gamma(S)}{\Gamma(S+n_{x_j})} \frac{\prod_k \Gamma(n_{c_k,x_j}+\alpha_k)}{\prod_k \Gamma(\alpha_k)}}{\frac{\Gamma(S)}{\Gamma(S+n^t)} \frac{\prod_k \Gamma(n_{c_k}+\alpha_k)}{\prod_k \Gamma(\alpha_k)}} > 1$$

- A Dirichlet prior distribution over the parameters is assumed with uniform alphas = S/|C|.
- **S** is considered the global sample size.

Figure 1: Example Iris Data Classification



- Firstly, the classification tree is induced following the classic recursive partitioning method for building CT. Each attribute is evaluated following the equation of Section 2.
- Let us see as there is no sample at the red bounded leaf. So there is no associated decision for that leaf.

- Secondly, for each node it is computed its associated weight accordingly to the red bounded quotient of Section 3.
- The weight of "Petal Width" is much higher than "Petal Length" because the partition of "Petal Width" is better than the partition of "Petal Length".

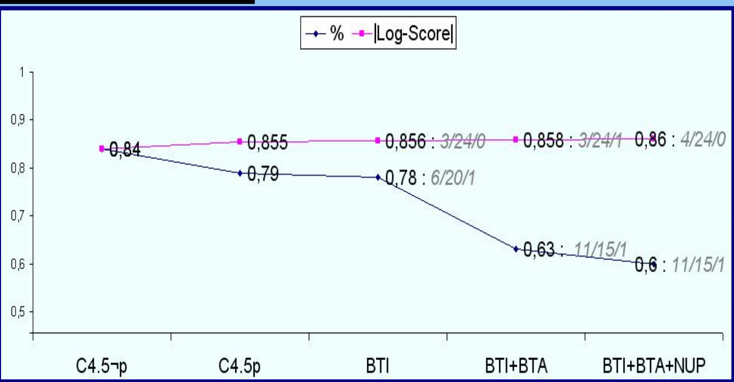
- Finally, the probabilities are weighted and updated following the summation equation of Section 3.
- As we can see, at the red bounded leaf there is now an associated decision. This effect would be the same than a post-pruning process, but the CPE are more precise with this approach.

3. Bayesian Tree Averaging (BMA)

- In many cases, branching by a node is only a little more probable than stopping the branching. So, there is uncertainty in this decision: Bayesian model averaging (BMA) [4] is an approach to deal with this uncertainty.
- Our application of BMA is an alternative of pruning the final tree. The probability at leaves are estimated as follows:

$$P(c_k|\mathbf{x}^{t_p}) \propto \sum_{i=1}^{p} \frac{n_{c_k\mathbf{x}^{t_i}}+\alpha_k}{n_{\mathbf{x}^{t_i}}+S} \prod_{j=1}^{i-1} \underbrace{\frac{P(M^{t_j}_{X^*}|\mathcal{D})}{P(M^{t_j}|\mathcal{D})}}_{W_i}$$

Figure 2: Results



- **%** → Percentage of correct classifications.
- **|Log-Score|** → Absolute Value of log-score. As lower it is as better the class probability estimates are.
- W/D/L → The number of databases where there is a statistically significant (at 1% level) improvement respect to the score (% or |Log-S|) of C4.5p (it is set as reference method).

4. Non-Uniform Priors (NUP)

- In previous analysis, uniform alpha values has been considered for Dirichlet prior distributions over the parameters.
- We test here a heuristic to define non-uniform alpha values.
- It is based on the fact that trees partition data and create subsets where there is no sample for some classes.

5. Experiments & Conclusions

- Methods were evaluated in 27 UCI data sets.
- We compare the following 5 methods:
  - C4.5 of Quinlan with (C4.5p) and without pruning (C4.5¬p).
  - BTI of Section 2, BTI+BTA of Section 3 and BTI + BMA + NUP.
  - Several **S** values were evaluated: S=1, S=2 and S=|C|.
- Two evaluated scores: the classic % of correct classification and the log-likelihood of the true class (log-Score), this last score is introduced with the aim of evaluate the quality of CPE.
- Results are presented in **Figure 2**: the mean value of both scores and the outputs of a corrected paired t-test are plotted. For simplicity, only models with S = 2 are showed.
- The main conclusions are:
  - BTI, BTA and NUP supposes an improvement in CPE and maintain the accuracy of C4.5p.
  - The Bayesian approach is a promise technique to deal with model uncertainty in CT.

References

[1] Pazzani et al. 1994. Reducing misclassification costs. In *International Conference of Machine Learning*, pages 217-225.

[2] Provost and Domingos. 2003. Tree induction for probability-based ranking. *Machine Learning*, 52(3):199-215.

[3] Heckerman, Geiger, and Chickering. 1994. Learning Bayesian networks: The combination of knowledge and statistical data. In *KDD Workshop*, pages 85-96.

[4] Hoeting, Madigan, Raftery and Volinsky. 1999. Bayesian model averaging: A tutorial. *Statistical Science*, 14(4):382-417.