Proceedings of the 7th International Conference on Statistics: Theory and Applications (ICSTA'25)
Paris, France - August, 2025
Paper No. 149
DOI: 10.11159/icsta25.149

Feature-Selective Oblique Trees for Regression: Application to STEM Graduate Wage Prediction in Italy

Andrea Carta[1], Luca Frigau[1]
[1]Department of Economics and Business Sciences, University of Cagliari,
Via Sant'Ignazio 17, Cagliari, Italy
andrea.carta88@unica.it; frigau@unica.it

Extended Abstract

Traditional decision trees are widely used but are limited by their axis-aligned splits, which can lead to large and complex models when handling high-dimensional or correlated data. Oblique decision trees attempt to overcome these issues by using linear combinations of predictors to build, at each node, the splitting hyperplane. Nevertheless, the majority of oblique tree methods are focused on classification tasks or use intensive computational optimization processes that prevent the interpretability and scalability of the trees [1]. In this work, we introduce a novel approach for constructing oblique decision trees for regression tasks. This method is called Selection variable weighted support vector machine Oblique Regression decision Tree (SORT) and addresses the limitations of traditional and oblique trees by integrating a variable selection process and a weighted support vector machine (SVM) [2] with a linear kernel into the decision tree framework.

At each node SORT selects the most correlated features with the target variable y, then transforms y into a dichotomous variable using the quantiles of its distribution. For each quantile, a weighted linear SVM is applied to find a splitting hyperplane, with the best one chosen based on deviance reduction. In particular, the weights assigned in the SVM to observations are computed as the absolute value of the scaled elements of y, ensuring that extreme values have a stronger influence on the splitting process. The process is repeated recursively until a stopping criterion is met.

We carried out a simulation study to assess SORT's performance under different data scenarios, such as noisy features, non-normal transformed variables, and the use of categorical variables. By analysing 3,840 simulated datasets, we found that selecting just two features at each split and using the median as the dichotomization threshold increases the tree's performance and computational efficiency. Moreover, this parametrization of SORT also increases the tree interpretability, as the resulting trees can be visualized and understood. In addition to this, SORT, consistently outperformed five other decision tree methods across the simulated datasets, including traditional CART and oblique trees such as ODT, CO2, HHCART, and BUTIA [3-7], in terms of predictive accuracy.

We also apply SORT to investigate a real-world problem, specifically focusing on the prediction of wages for graduates in Science, Technology, Engineering and Mathematics (STEM) fields in Italy. The results indicate that SORT outperforms other oblique tree methods. Furthermore, using only two predictors at each node, there is an increase in interpretability, allowing us to determine the main factors impacting salary levels, potentially revealing structural issues such as the gender wage gap and Italy's North-South divide.

In conclusion, SORT offers a powerful and interpretable alternative to traditional regression trees, particularly in settings with complex feature interactions. Future work may investigate non-linear kernels within the weighted SVM framework and ensemble versions of SORT to further improve flexibility and accuracy.

Acknowledgments:
This study was funded by the European Union - NextGenerationEU, in the framework of the GRINS -Growing Resilient, INclusive and Sustainable project (CUP F53C22000760007). The views and opinions expressed are solely those of the authors and do not necessarily reflect those of the European Union, nor can the European Union be held responsible for them.

References

[1] S. Chaturvedi and S. Patil, “Oblique decision tree learning approaches—a critical review,” *Int. J. Comput. Appl.*, vol. 82, no. 13, 2013.

[2] X. Yang, Q. Song, and Y. Wang, “A weighted support vector machine for data classification,” *Int. J. Pattern Recognit. Artif. Intell.*, vol. 21, no. 5, pp. 961–976, 2007.

[3] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*. New York: Chapman & Hall, 1984.

[4] H. Zhan, Y. Liu, and Y. Xia, “Consistency of the oblique decision tree and its random forest,” *arXiv preprint*, arXiv:2211.12653, 2022.

[5] M. Norouzi, M. D. Collins, D. J. Fleet, and P. Kohli, “Co2 forest: Improved random forest by continuous optimization of oblique splits,” *arXiv preprint*, arXiv:1506.06155, 2015.

[6] D. C. Wickramarachchi, B. L. Robertson, M. Reale, C. J. Price, and J. Brown, “Hhcart: An oblique decision tree,” *Comput. Stat. Data Anal.*, vol. 96, pp. 12–23, 2016.

[7] R. C. Barros, P. A. Jaskowiak, R. Cerri, and A. C. Carvalho, “A framework for bottom-up induction of oblique decision trees,” *Neurocomputing*, vol. 135, pp. 3–12, 2014.