Summary of the work entitled
The tree based linear regression model for hierarchical categorical variables

1 Summary

Many real-life applications consider nominal categorical predictor variables that have a hierarchical structure, e.g. economic activity data in Official Statistics. In this paper, we focus on linear regression models built in the presence of this type of nominal categorical predictor variables, and study the consolidation of their categories to have a better tradeoff between interpretability and fit of the model to the data. We propose the so-called Tree based Linear Regression (TLR) model that optimizes both the accuracy of the reduced linear regression model and its complexity, measured as a cost function of the level of granularity of the representation of the hierarchical categorical variables. We show that finding non-dominated outcomes for this problem boils down to solving Mixed Integer Convex Quadratic Problems with Linear Constraints, and small to medium size instances can be tackled using off-the-shelf solvers. We illustrate our approach in two real-world datasets, as well as a synthetic one, where our methodology finds a much less complex model with a very mild worsening of the accuracy.

2 State-of-the-art

Categorical variables are increasingly present in a number of real-world applications. For example, in the healthcare field, data may contain high-cardinality categorical variables describing diagnoses and prescriptions [22]. They may also appear in social and economic studies [31] or in Natural Language Processing [30], to name a few. Interpreting and visualizing information extracted from complex data is at the core of Data Science [4, 11, 24, 29, 35], and this is also the case for categorical variables where the information may be disaggregated across many categories. Mathematical Optimization is an important tool to build, in an efficient manner, data analysis models that can achieve a high accuracy [9, 16, 18], while being able to incorporate desirable properties, such as being parsimonious [2, 3, 5, 6, 7, 12, 27], or tackling multiple objectives, such as the *bias-variance tradeoff* [21].

In the linear regression setting, to enhance the interpretability of the model and reduce the risk of overfitting in the presence of high-cardinality categorical variables, some works have fused categories, i.e., they are forced to share the same estimated coefficient, see [10, 33] and references therein.

3 Contribution to the literature

In this paper, we are interested in the fusion of categories for a special case, those variables that have a hierarchical structure in their categories. This kind of variable appears in different fields of research, such as nested spatial data in Spatial Statistics [19], as for example the European Union with the NUTS classification (nomenclature of territorial units for statistics), where the small regions for specific diagnoses are consolidated at basic regions for the application of regional policies and these, in turn, are consolidated at major socio-economic regions. They also appear in behavioral data in Retail

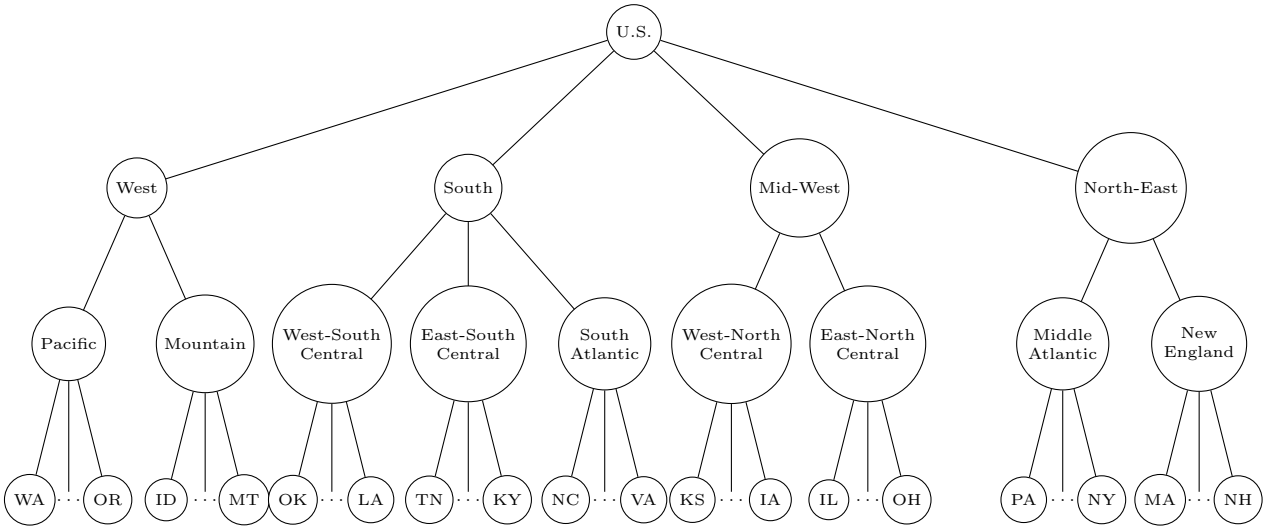

Figure 1: Tree representation of the variable *geography* in the `cancer-reg` dataset

Business Analytics [20], since each retailer chain maintains a product hierarchy, which is necessary to conduct business processes such as store replenishment. Economic activity data in Official Statistics [17, 23] is another example of hierarchical categorical variable, where the interdependency of activities forms a hierarchy. Thus, in this paper, we study the mathematical optimization problem that trades off, in linear regression models, accuracy and model complexity, while exploiting the structure of the nominal hierarchical categorical variables.

The categories of hierarchical categorical variables can be arranged as a directed tree, i.e., a directed graph with a root node and a unique path from each node to the root. As an example, consider the real-world dataset `cancer-reg` [32] used in the numerical section, with individuals from the United States of America (U.S.), where variable *geography* is a categorical variable with a hierarchical structure. See Figure 1 for the tree associated with its categories. The leaf nodes correspond to the states, going upstream we find the subregions and then the regions, which, in turn, are directly connected with the root node. The question arises as to whether the highest level of granularity (states) is necessary to build the linear regression model, or whether we can merge categories at the bottom of the tree into a broader category upstream in the tree.

Reducing the granularity of the representation of hierarchical categorical variables has several advantages. First, it is a step towards enhancing the interpretability of the linear regression model, where fewer coefficients need to be estimated and interpreted [14, 10]. Second, if the samples of individuals associated with categories are homogeneous enough, a very granular representation would yield an overparametrized model. Instead, we could merge these categories into a broader one upstream the tree, thus having more observations to estimate fewer coefficients. The homogeneity together with the increase in sample size ensure lower errors in the estimation of the coefficients of the broader categories [25]. Third, and again if the samples of individuals associated with categories are homogeneous enough, a very granular representation will yield higher data gathering costs [13, 34], if, for instance, the surveying costs are asymmetric. Indeed, we would need to ensure a large enough sample for each category in the representation, even though the cost of surveying may be high for some of these categories. By merging homogeneous categories into a broader one upstream the tree, we can sample from a larger subpopulation lowering these data gathering costs. Fourth, our methodology

can identify where j is an irrelevant predictor [5, 8, 15] by consolidating individuals at the root node $r(\mathcal{T}_j)$. Finally, the consolidation of information is important when having data privacy considerations, [26, 28], since it is well-known that more detailed information is linked to confidentiality concerns [1].

References

[1] D. Baena, J. Castro, and A. Frangioni. Stabilized Benders Methods for Large-Scale Combinatorial Optimization, with Application to Data Privacy. *Management Science*, 66(7):3051–3068, 2020.

[2] S. Benítez-Peña, R. Blanquero, E. Carrizosa, and P. Ramírez-Cobo. Cost-sensitive Feature Selection for Support Vector Machines. *Computers & Operations Research*, 106:169 – 178, 2019.

[3] D. Bertsimas and A. King. OR Forum–An Algorithmic Approach to Linear Regression. *Operations Research*, 64(1):2–16, 2016.

[4] D. Bertsimas, A. O'Hair, S. Relyea, and J. Silberholz. An Analytics Approach to Designing Combination Chemotherapy Regimens for Cancer. *Management Science*, 62(5):1511–1531, 2016.

[5] D. Bertsimas, J. Pauphilet, and B. V. Parys. Sparse Regression: Scalable Algorithms and Empirical Performance. *Statistical Science*, 35(4):555 – 578, 2020.

[6] D. Bertsimas and B. Van Parys. Sparse high-dimensional regression: Exact scalable algorithms and phase transitions. *Annals of Statistics*, 48(1):300–323, 2020.

[7] R. Blanquero, E. Carrizosa, A. Jiménez-Cordero, and B. Martín-Barragán. Variable selection in classification for multivariate functional data. *Information Sciences*, 481:445 – 462, 2019.

[8] R. Blanquero, E. Carrizosa, C. Molero-Río, and D. Romero Morales. Sparsity in optimal randomized classification trees. *European Journal of Operational Research*, 284(1):255 – 272, 2020.

[9] L. Bottou, F. Curtis, and J. Nocedal. Optimization Methods for Large-Scale Machine Learning. *SIAM Review*, 60(2):223–311, 2018.

[10] E. Carrizosa, M. Galvis Restrepo, and D. Romero Morales. On clustering categories of categorical predictors in generalized linear models. *Expert Systems with Applications*, 182:115245, 2021.

[11] E. Carrizosa, V. Guerrero, and D. Romero Morales. Visualizing data as objects by DC (difference of convex) optimization. *Mathematical Programming, Series B*, 169:119–140, 2018.

[12] E. Carrizosa, V. Guerrero, and D. Romero Morales. On mathematical optimization for clustering categories in contingency tables. Technical report, Universidad Carlos III, Madrid, Spain, `https://www.researchgate.net/publication/341079651_On_mathematical_optimization_for_clustering_categories_in_contingency_tables`, 2020.

[13] E. Carrizosa, B. Martín-Barragán, and D. Romero Morales. Multi-group support vector machines with measurement costs: A biobjective approach. *Discrete Applied Mathematics*, 156:950–966, 2008.

[14] E. Carrizosa, A. Nogales-Gómez, and D. Romero Morales. Clustering categories in support vector machines. *Omega*, 66:28 – 37, 2017.

[15] E. Carrizosa, A. V. Olivares-Nadal, and P. Ramírez-Cobo. A sparsity-controlled vector autoregressive model. *Biostatistics*, 18(2):244–259, 2016.

[16] E. Carrizosa and D. Romero Morales. Supervised classification and mathematical optimization. *Computers and Operations Research*, 40(1):150–165, 2013.

[17] European Commission. *NACE Rev. 2 – Statistical classification of economic activites in the European Community*. Luxembourg: Office for Official Publications of the European Communities, 2008. https://ec.europa.eu/eurostat/documents/3859598/5902521/KS-RA-07-015-EN.PDF.

[18] K. Fountoulakis and J. Gondzio. A second-order method for strongly convex ℓ_1-regularization problems. *Mathematical Programming*, 156(1):189–219, 2016.

[19] C. A. Gotway and L. J. Young. Combining incompatible spatial data. *Journal of the American Statistical Association*, 97(458):632–648, 2002.

[20] A. Griva, C. Bardaki, K. Pramatari, and D. Papakiriakopoulos. Retail business analytics: Customer visit segmentation using market basket data. *Expert Systems with Applications*, 100:1 – 16, 2018.

[21] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Science & Business Media, 2009.

[22] P. B. Jensen, L. Jensen, and S. Brunak. Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics*, 13:395–405, 2012.

[23] T. Katz-Gerro and J. López Sintas. Mapping circular economy activities in the European Union: Patterns of implementation and their correlates in small and medium-sized enterprises. *Business Strategy and the Environment*, 28(4):485–496, 2019.

[24] J. Kleinberg, H. Lakkaraju, J. Leskovec, J. Ludwig, and S. Mullainathan. Human Decisions and Machine Predictions. *The Quarterly Journal of Economics*, 133(1):237–293, 2017.

[25] M. LeBlanc and R. Tibshirani. Monotone shrinkage of trees. *Journal of Computational and Graphical Statistics*, 7(4):417–433, 1998.

[26] X.-B. Li and S. Sarkar. Against classification attacks: A decision tree pruning approach to privacy protection in data mining. *Operations Research*, 57(6):1496–1509, 2009.

[27] J. Lin, C. Zhong, D. Hu, C. Rudin, and M. Seltzer. Generalized and Scalable Optimal Sparse Decision Trees. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 6150–6160. PMLR, 2020.

[28] R. Lu, H. Zhu, X. Liu, J. K. Liu, and J. Shao. Toward efficient and privacy-preserving computing in big data era. *IEEE Network*, 28(4):46–50, 2014.

[29] D. Martens, B. Baesens, T. V. Gestel, and J. Vanthienen. Comprehensible credit scoring models using rule extraction from support vector machines. *European Journal of Operational Research*, 183(3):1466 – 1476, 2007.

[30] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*, 2013.

[31] D. Pauger and H. Wagner. Bayesian Effect Fusion for Categorical Predictors. *Bayesian Analysis*, 14(2):341 – 369, 2019.

[32] N. Rippner. Cancer Trials, 2017. Retrieved from `http://data.world/exercises/linear-regression-exercise-1/workspace/file?filename=cancer_reg.csv`.

[33] B. G. Stokell, R. D. Shah, and R. J. Tibshirani. Modelling high-dimensional categorical data using nonconvex fusion penalties. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 83(3):579–611, 2021.

[34] P. Turney. Cost-sensitive classification: Empirical evaluation of a hybrid genetic decision tree induction algorithm. *Journal of Artificial Intelligence Research*, 2:369–409, 1995.

[35] B. Ustun and C. Rudin. Supersparse linear integer models for optimized medical scoring systems. *Machine Learning*, 102(3):349–391, 2016.