TACKLING SIMPSON'S PARADOX IN BIG DATA USING CLASSIFICATION & REGRESSION TREES

Research in Progress

Shmueli, Galit, Indian School of Business, Hyderabad, India, galit_shmueli@isb.edu

Yahav, Inbal, Bar Ilan University, Ramat Gan, Israel, inbal.yahav@biu.ac.il

Abstract

This work is aimed at finding potential Simpson's paradoxes in Big Data. Simpson's paradox (SP) arises when choosing the level of data aggregation for causal inference. It describes the phenomenon where the direction of a cause on an effect is reversed when examining the aggregate vs. disaggregates of a sample or population. The practical decision making dilemma that SP raises is which level of data aggregation presents the right answer.

We propose a tree-based approach for detecting SP in data. Classification and regression trees are popular predictive algorithms that capture relationships between an outcome and set of inputs. They are used for record-level predictions and for variable selection. We introduce a novel usage for a cause-and-effect scenario with potential confounding variables. A tree is used to capture the relationship between the effect and the set of cause and potential confounders. We show that the tree structure determines whether a paradox is possible. The resulting tree graphically displays potential confounders and the confounding direction, allowing researchers or decision makers identify potential SPs to be further investigated with a causal toolkit.. We illustrate our SP detection approach using real data for both a single confounder and for multiple confounder in a large dataset on Kidney transplant waiting time.

Keywords: Simpson's Paradox, CART, Big Data, Casual Effect, Classification and Regression Trees

1 Introduction

With the growing availability of data at more granular levels, decision making has expanded from aggregate-level to personalized decisions. In medicine, we see a shift towards personalized medicine. In marketing, personalized offers and customer experiences are now common. Data-based decision making is important at different aggregation levels: population aggregates are needed for overall policy making (“macro decisioning"), while micro-data are used for personalized decisioning (“micro decisioning”. An important question is therefore, *what data are needed for what level of decision making*?. This question relates to the concept of Information Quality (Kenett and Shmueli, 2014), which is the potential of a dataset to achieve a specific (scientific or practical) goal using a given empirical analysis method. One dimension of Information Quality is *data resolution*, which refers to the measurement scale and aggregation level of the data. This dimension is the focus of this paper. In particular, we ask which data aggregation level is needed in the context of macro decisioning.

A perplexing issue in the context of choosing the level of aggregation for macro decisioning is Simpson's paradox (Simpson, 1951). The paradox describes the phenomenon where the direction of a cause on an effect appears reversed when examining the aggregate vs. disaggregates of a sample or a population. The practical macro decision making question that Simpson's paradox raises is which level of data aggregation presents the results of interest. This practical question raises the challenge of identifying potential confounders and then establishing a criterion for deciding if and which of the potential confounders should influence the decision making.

One might think that with sufficiently large samples, it is always safer to use the disaggregate data, which are potentially more homogeneous. While this might be true for micro decisioning (personalized decision making), it is not necessarily the case for macro decisioning, where the goal is to evaluate an overall effect. Pearl shows that, in many cases it is the aggregated rather than the partitioned data that gives the correct choice of action.

Pearl describes Simpson's paradox as “the phenomenon whereby an event *C* increases the probability of *E* in a given population *p* at the same time, decreases the probability of *E* in every subpopulation of *p*". Pearl warns that Simpson's paradox can only be resolved when the observational data (frequency tables) are combined with a causal theory. He shows that the same data table can result from different causal paths and therefore the underlying causal structure is unidentifiable from the data alone. In other words, once the effect, cause, and potential confounding variable are singled out, a causal narrative is required for determining which level of aggregation to use for decision making.

The literature on Simpson's paradox has focused on explaining the phenomenon, quantifying its magnitude (Blyth, 1972, and Zidek, 1984), the conditions where it disappears (Bishop et al, 2007) and its frequency (Pavlides and Perlman, 2009). Explanations range from mathematical arguments (*Simpson's Reversal of Inequalities* looks at the paradox algebraically in terms of ratios of numbers; see Stanford Encyclopedia of Philosophy) to ‘group inhomogeneity' explanations (subpopulations of different sample sizes) to ‘confounding factor' explanations. Yet, there still remains the question of which result to act on in the presence of contradicting directions. As Schield (1999) stated, “even if Simpson's Paradox were readily understood, it is not easily anticipated. There is no test for determining whether an association is spurious". Pearl (1999) notes that “there is no statistical criterion that would warn the investigator against drawing the wrong conclusion or would indicate which table represents the correct answer".

The focus of this work is the practical issue of identifying potential Simpson’s paradoxes in Big Data. Bid Data are now common in many areas and are the basis for important policy decisions. Our focus is on the pre-paradox scenario, where the goal is to identify potential confounding variables among a broad set of variables. The suspect confounders are those that would lead to Simpson's paradox. Once the potential confounders are identified, the researcher can investigate plausible causal narratives (such as self-selection) to determine the adequate data aggregate level to use.

We take a data-driven approach that searches the terrain of possible relationships between the outcome of interest and the set of cause and potential confounders. A tree-based approach is applied to micro-level data and automatically identifies existing relationships and their structure. The result graphically displays potential confounders and the structure of confounding, allowing the researcher or decision maker to identify potential Simpson's paradox relationships to be further investigated with a causal toolkit such as Pearl's “back-door" test.

2 Classification & Regression Trees for Causal Modelling

2.1 Standard Use: Trees in Predictive Modelling

Classification and Regression Trees (“trees”) are a popular machine learning algorithm for predicting an outcome from a set of predictors. Trees use the notion of recursive partitioning, where at each step the sample is partitioned into subsamples to create the most homogeneous subsamples in terms of the outcome. Partitioning is obtained by searching across all the predictors and all their values for the best split. The best split is the one that splits the data into the most homogeneous sub-samples in terms of the outcome variable. The result of a single step is therefore a split on one predictor at a particular value (or in the case of a categorical predictor, a split into groups of categories). Given a partition from one step, the algorithm seeks the next (conditional) split. The end result is a set of rules based on predictor thresholding (“if Age<20 and Employed=Yes”) that splits the data into non-overlapping subsamples. Variants of trees exist in terms of choosing splitting criteria (Gini index, entropy, and statistical significance of independence tests) and approaches to avoiding over-fitting (stopping tree growth, pruning, etc.). The resulting tree is then used for predicting the outcome of new records.

Similar to Stepwise regression techniques, trees are also a useful tool for variable selection and dimension reduction. Important predictors are those chosen by the algorithm for splitting the sample into homogeneous subsamples. The presence or absence of a predictor from a tree is therefore informative of its importance. Due to the recursive nature of the algorithm, predictors that appear at the top of the tree (those who generated early splits) are more likely to have stronger predictive power.

An advantage of trees over other predictive algorithms is that a tree is interpretable in terms of the relationship between the predictors and the outcome, and can be translated to a set of easily understood IF-THEN rules.

2.2 Non-Standard Use: Trees for Causal Modelling

While trees are common in predictive analytics, they are nearly absent from causal modelling. In predictive modelling, trees are used for micro-decisioning, where a prediction is required for each new record. We introduce an approach for using trees in macro decisioning in the context of explanatory modelling. In particular, we use trees to automatically identify potential confounding variables in a high-dimensional dataset, which would cause Simpson's paradox.

Our use of trees in the context of explanatory modelling and in particular Simpson's paradox differs from predictive modelling in a few ways. First, we use full grown trees (“full trees") which deliberately overfit the sample. Second, we use conditional-inference trees (Hothorn et al. (2006)) where variable choice and splitting values are based on statistical tests of independence rather than on cross-validation or holdout data pruning. Second, we are interested in the tree structure itself: not only which predictors are present, but also which predictors are absent, and what is the ordering of the splits. Unlike predictive modelling, we do not use the tree to predict (“score") new records.

Consider a dataset with an outcome (effect) of interest (Y), a cause variable (X) and a set of potential confounding variables (Z). We fit a tree with the outcome Y and predictors that include both X and Z. If Y is categorical, we fit a classification tree. If Y is numerical, we fit a regression tree. The structure of the resulting tree will then yield the potential confounders and their confounding behaviour. In particular, we examine the presence and absence of each of X and Z variables as splitting variables in the tree, as well as the splitting sequence. In the next Section we describe two different types of trees. The first type, full trees, reflects the full contingency table data in the same format used for displaying Simpson's paradox. The second type of trees, conditional-inference trees, incorporates sampling error by using statistical inference for testing the significance of the relationships. The conditional-inference trees are typically smaller than the full trees, in terms of having a smaller number of splits (the trees contain only significant splits).

3 Simpson's Paradox and Trees

Simpson's paradox is classically displayed using contingency tables, where rows and columns are used for conditioning on X and Z (or vice-versa) and the cell values are counts, probabilities, percentages or numerical summaries such as averages of Y. The table then allows comparing the conditional values for different levels of Z, thereby conditioning on Z. The same information can be clearly displayed in a tree of Y on predictors X and Z.

3.1 Tree Structure and Simpson's Paradox

If we consider X and a single confounder Z, there are potentially five types of full-grown trees (see Figure 1):

Type 1 Tree: no splits ("stub"). A stub indicates no association between Y and X, nor between Y and Z. Hence, splitting by either does not create more homogeneous subsamples.

Type 2 Tree: only splits on X. This structure indicates an association between X and Y, with no confounding effect of (the absent) Z.

Type 3 Tree: only splits on Z. The absence of X indicates no association between X and Y. The presence of Z indicates that Y is associated with Z.

Type 4 Tree: split first on Z then on X. This tree indicates a relationship between X and Y that is confounded by Z.

Type 5 Tree: split first on X then on Z.

Trees of types 1, 2, and 3 exclude X and/or Z as splits and therefore the corresponding contingency tables would not exhibit Simpson's paradox. The type 5 tree will also correspond to a no-paradox contingency table. The reason is that the ordering of splits, where X occurs earlier, indicates that the X-Y relationship is stronger than the Z-Y relationship. According to Cornfield's condition, Simpson's paradox will only occur if Z has the strength - the effect size - necessary to nullify or reverse an observed association between X and Y (Schield, 1999).

Among the five tree types, therefore, only the type 4 tree corresponds to a contingency table that can exhibit Simpson's paradox. The ordering of the splits with Z earlier than X assures that the reversal is possible according to Cornfield's condition. Note that a type 4 tree does not guarantee a paradox. A type 4 tree can also correspond to a contingency table exhibiting a paradox that is statistically insignificant, a partial paradox where the effect is reversed only for some of the subgroups, or no paradox, where the effect for different subgroups differs in magnitude but not direction.

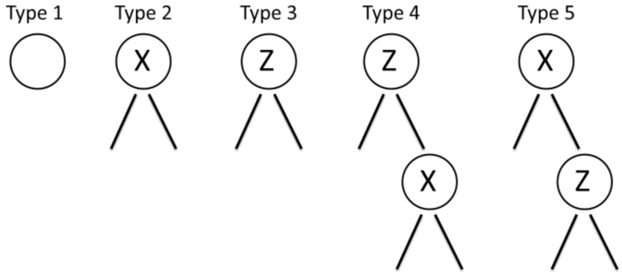

Figure 1. The five types of tree structures for a single confounding variable.

3.2 Impurity-based vs. statistical-inference-based trees

Different tree algorithms differ in terms of the criteria for variable selection and splitting rules. Classic predictive algorithms such as CART and C5.0 use "node impurity" (the homogeneity of Y) measures to select variables and splits for generating the full tree (which is then followed by a pruning step). In contrast, CHAID and conditional-inference (CI) trees (Hothorn et al., 2006) use statistical tests of association of the predictors with Y. In this scheme of recursive partitioning, at each stage predictors are ordered according to the significance of their (conditional) association with Y, and the strongest associated predictor is chosen for splitting. However, the split is only performed if the significance level exceeds a pre-determined threshold. Tree growth is stopped when the significance threshold is not met, thereby implying insufficient association strength. The result regarding tree structure and Simpson's Paradox is general to full trees grown using either type of algorithm. In the context of a single confounding variable, trees based on statistical tests of association have the advantage of providing information on the statistical significance of the Paradox. Setting the significance threshold to a preferred value will therefore limit findings to paradoxes that are significant at least at that level. For multiple confounders, the significance threshold imposed on a conditional-inference tree is no longer aligned with a tree that has multiple layers of splits. We thus introduce the X-Terminal tree.

4 Single Confounder: Conditional-Inference Tree

We use the example of death sentence rates from Agresti (2012). The data include information on 326 murder cases. In each case, data is available on the race of the defendant (X={white,black}), whether the outcome was a death sentence (Y={yes,no}), and the race of the victim (Z={white,black}).

The question of interest is whether the defendant's race affects the probability for a death sentence. The potential confounder is the race of the victim. Examining the contingency table of the aggregate and disaggregate data (Table 1) indicates Simpson's paradox. The aggregate rates indicate that white defendants are more likely to get the death sentence than black defendants (11.88% vs 10.24%). In contrast, disaggregation by victim race indicates that white defendants are less likely to get the death sentence, when the victim is black (0% vs 5.83%) as well as when the victim is white (12.58% vs 17.46%).

Figure 2 shows the full-grown and CI classification trees of death on predictors X (defendant's race) and Z (victim's race). As expected, the resulting full tree is of type 4, with an initial split on Z followed by splits on X. The tree also clearly displays a higher death sentence rate for black defendants (bar charts for nodes 3,5) compared to white defendants (nodes 4,6). The CI tree indicates that the Simpson's paradox observed in the contingency table (and full-grown tree) is not generalizable to the population, when considering statistical significance. That is, the death sentence rate is unrelated to the race of the defendant.

Defendant's Race	Victim's Race	Death Sentence Rate	
Black	Black	5.83%	10.24%
	White	17.46%	
White	Black	0.00%	11.88%
	White	12.58%	

Table 1. Death sentence rate, by defendant's race and victim's race

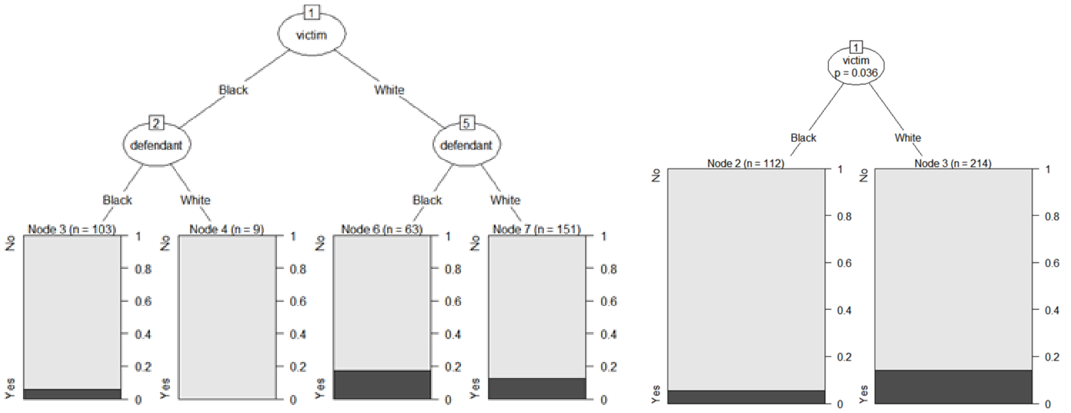

Figure 2. Full-grown classification tree (left) and CI tree (right) based on the death sentence disaggregate data, with defendant and victim race as predictors.

5 Multiple Confounders: X-Terminal Tree

In data with a large number of potential confounders, the resulting full tree might be overwhelming in size. For detecting confounders that might lead to Simpson's paradox, we have a sequential process: First detect tree type. If not type 4, then no paradox and stop here. If type 4, then examine only tree paths that lead to a split on X. This constrained space of information can be achieved by growing trees so that when an X split is encountered that branch is no longer grown. Using this process, the resulting tree will either be a non-paradox full-grown tree (of type 1,2,3 or 5), or else it will be a type 4 tree with at least one terminal node at X.

We illustrate the application of our tree approach to a large dataset (*n*=104,000) with multiple potential confounders (*p*=19). The question of interest is the effect of waitlist kidney transplant patients' race on their waiting time for transplant. A naïve comparison of the average wait time of black patients (717 days) and white patients (473 days) indicates a statistically and practically significant difference. The waitlist dataset includes many variables with patient-level information that are potential confounders. In this example, the outcome of interest is continuous and therefore we use a regression X-terminal tree. The resulting X-terminal tree (Figure 3) is of type 4, with Race appearing as a split below a few confouders. Examining all paths terminating with Race, none of them exhibit Simpson's paradox. We therefore conclude that there is no paradox in the dataset, and that black recipients wait longer for kidney transplants compared to white recipients in all the subgroups examined.

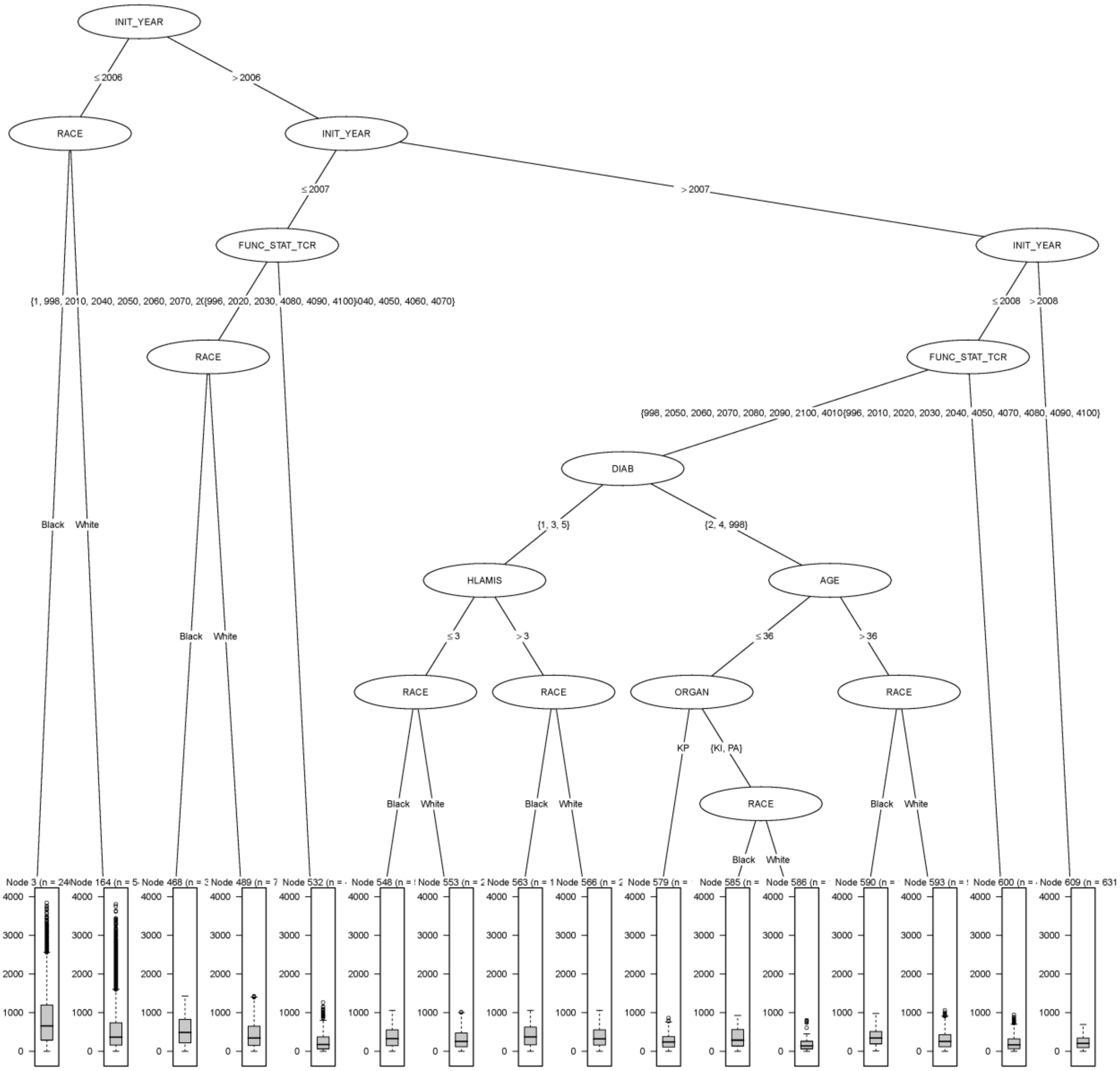

Figure 3: X-Terminal regression tree for effect of race on wait time for transplant. Boxplots in terminal nodes show wait time distribution.

6 Discussion

We have introduced the use of a popular predictive algorithm, classification and regression trees, to a context where it is rarely used: explanatory modelling. Using trees in this very different context warrants several conceptual differences in its use. First, we are mostly interested in the presence or absence of *X* and *Z* as splitting variables in the tree. In predictive modelling, the main interest is in the terminal nodes, which provide the prediction value. The absence of presence of predictors in the tree is useful as a secondary goal of variable selection.

Second, in predictive modelling the main concern is that of over-fitting the tree to the training data, so that it is unable to generalize to new records, thereby having low predictive power. Hence, there exist various approaches to avoiding over-fitting, the most common approach by pruning the tree using a holdout dataset or cross-validation. In the explanatory context, over-fitting is less of a concern. Hence, the choice of using a full tree or a pruned tree is approached with a different mind-set. In our context, we chose to use trees with stopping rules but with no pruning.

Third, the tree approach has a key advantage over using a logit model with X and Z and an interaction term. While an insignificant interaction term would rule out a paradox, a significant interaction term would not distinguish between a paradox and non-paradox because it would not tell us which of the X-Y or Z-Y relationships are stronger. This is the advantage of the tree approach.

Our result regarding tree structure applies not only to a single confounder *Z*, but to multiple confounding variables. In this case, the ordering of the splits in full grown trees is the key to detecting a possible paradox. Simpson's paradox can only occur when a confounder *Z* is the top split and *X* is a split that appears somewhere in the tree (in our notation, tree type 4).

With multiple potential confounding variables, the advantage of using full-grown trees over scanning all possible contingency tables is clear: the tree approach easily scales up using automated variable selection, thereby detecting confounding variables that might display Simpson's paradox. Another advantage of the tree over contingency tables is that it easily generalizes to variables *X* and *Z* that are not categorical. In such cases, creating a contingency table requires binning the continuous variables, yet how to bin those variables is unclear. In contrast, the tree will automatically choose the best binning in terms of the strongest *Z-Y* association and if that association exists, it will be displayed in the tree as splits based on those bins.

One challenge with scaling up the tree approach to multiple confounders is in terms of statistical significance, where the significance threshold imposed on a conditional-inference tree is no longer aligned with a tree that has multiple layers of splits. Another challenge is determining the structure of a very large tree. Our approach is to modify the tree stopping criterion to depend on the tree type. We will expand on this approach in our talk and show initial results.

References

Kenett, R.S. and Shmueli, G. (2014). On Information Quality. Journal of the Royal Statistical Society, Series A, 177(1), 3-27.

Simpson, E.H. (1951). The Interpretation of Interaction in Contingency Tables. Journal of the Royal Statistical Society B, 13, 238-241.

Judea Pearl (2009). Causality: Models, Reasoning, and Inference. Cambridge University. ISBN 0-521-77362-8.

Blyth, C.R. (1972). On Simpson's paradox and the sure-thing principle. J. American Statistical Association, 67, 364-366.

Zidek J. (1981). Maximal Simpson-disaggregations of 2×2 tables. Biometrika, 71(1), 187-190.

Bishop Y.M., Fienberg, S.F. and Holland, P.W. (2007). Discrete Multivariate Analysis. Springer Verlag. ISBN: 0-387-72805-8.

Pavlides, M.G. and Perlman, M D. (2009). How Likely is Simpson's Paradox? The American Statistician, 63(3), 226-233.

Schield, M. (1999). Simpson's Paradox and Conr_eld's Conditions", in Proceedings of the Section on Statistical Education. Joint Statistical Meeting of the American Statistical Association, 106-111.

Hothorn, T., Hornik, K. and Zeileis, A. (2006). Unbiased Recursive-Partitioning: A Conditional Inference Framework. Journal of Computational and Graphical Statistics, 15(3), 651-674.

Agresti, A. (2012). Categorical Data Analysis, Third Edition. Wiley and Sons.