Biostatistics Research
https://ojs.wiserpub.com/index.php/BSR/

UNIVERSAL WISER PUBLISHER

Research Article

A Simulation Study Comparing Tree-Based Methods in Identifying Interactions of Continuous and Binary Variables for Prediction of Increased Risk of Disease

Sybil Prince Nelson

Washington and Lee University, 204 W Washington St., Lexington, VA 24450, United States
E-mail: sprincenelson@wlu.edu

Received: 8 November 2022; **Revised:** 31 March 2023; **Accepted:** 3 April 2023

Abstract: Tree-based methods are commonly used to create models that predict an output based on several input variables. Classification and Regression Trees (CARTs) is a popular algorithm that builds tree-like graphs for predicting continuous and categorical dependent variables, but it has been shown to be biased toward the inclusion of continuous variables. Conditional inference is a technique used to alleviate this bias. C.Logic is an alternative tree-based method that uses Boolean logic to create classification trees. Previous research has shown that C.Logic is superior to CART in identifying interactions that lead to an increased risk of disease. No comparison has been made between the C.Logic package and CART with conditional inference as found in a package called Party. In this paper, a simulation study is used to compare the capability of these two algorithms to identify interactions between continuous and binary variables. It is found that while both methods succeed in identifying correct interactions, C.Logic is more effective. The C.Logic algorithm does a better job of alleviating the bias toward continuous variables when attempting to identify interacting variables that lead to an increased risk of disease.

***Keywords*:** dichotomization, biostatistics, tree graphs, conditional inference, interactions, classification and regression trees, logic regression, optimal cutpoints

1. Introduction

In medical practice, it is often necessary to dichotomize a variable perhaps for patient care or diagnosis [1]. For example, a patient with a total cholesterol level of above 239 mg/dL might be put on medication. Thus, the continuous variable of cholesterol level is dichotomized and made to be binary. Also, if the patient's high cholesterol is interacting with other factors, this may lead to a greater risk of certain diseases. Research shows that complex diseases may be influenced by the interactions between several clinical, environmental, and genetic variables [2-6]. If the interaction of variables is not considered, or if continuous variables are prioritized over binary variables increased risk of disease may fail to be detected [5, 6].

In a statistical interaction, the association of an effect measure (e.g. age) with outcome differs in the presence of a third variable (e.g. cholesterol level). Thus, for example, the association between age and disease may not be detected unless it is paired with another factor such as cholesterol level. This type of interaction may be difficult to detect with traditional statistical applications. Also, it is possible that only the interaction between age and high cholesterol leads to

an increased risk of disease and not high cholesterol alone or age alone.

Logistic regression is a common statistical approach often used to model dichotomous outcomes such as disease or no disease. When investigating interactions, however, they should be hypothesized *a priori*. When there are a small number of variables, all possible combinations of the variables can be included in the model without difficulty. For example, if there are 4 main effects, the number of total terms to include in the model would be $2^4 - 1 = 15$. If the number of variables increases to 20, however, the number of model terms becomes $2^{20} - 1 = 1{,}048{,}575$ which is infeasible to include in a single model. When the number of parameters exceeds the number of observations, which is common in genetic data, then logistic regression will fail. Further, the interaction of two variables can not be included in the model unless each of the main effects is also included.

Decision trees are non-parametric and semi-parametric classification models that result in easily interpreted tree-like graphs. They can easily identify interactions associated with increased risk of disease even in data sets with large numbers of variables which would be difficult with logistic regression. In decision trees, possibly significant variables do not have to be identified *a priori*. All variables can be considered. Classification and Regression Trees (CARTs), a common decision tree method, uses recursive partitioning to create homogeneous rectangular subsets in the data. Figure 1 gives an example of a simple CART built from three binary predictors. This tree predicts an individual to be in category 0 if $X_1 = 0$ and $X_2 = 0$ or if $X_1 = 0$, $X_2 = 1$, and $X_3 = 0$. It predicts category 1 if $X_1 = 1$, or if $X_1 = 0$, $X_2 = 1$ and $X_3 = 1$.

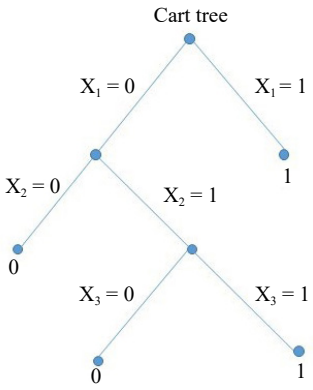

Figure 1. An example of a CART that predicts category 0 or category 1

Though decision trees, specifically CART, address the limitations of determining variables *a priori* and the number of parameters, there is one more concern. In our example, age is a continuous variable while smoking is binary or perhaps categorical. CART is biased toward the inclusion of continuous variables meaning that if it has a choice of including a continuous variable or a binary variable in the model, it will choose the continuous [7]. Conditional inference is used in the Party package to offset this bias. The conditional distribution is utilized to measure the association between response and potential variables which is then used as the basis for the unbiased selection of variables to be added to the model [8]. Conditional inference trees use a significance test, or permutation test, to input variables into the model whereas general recursive partitioning maximizes an information measure selecting the variable showing the best split [8]. This results in a selection bias towards variables with many possible splits [8].

Logic regression is an alternative tree-based method that uses Boolean logic to model a binary outcome and it is especially effective in finding interactions of variables [9]. The use of Boolean logic makes logic regression inherently more flexible than CART. This is because while all CARTs can be written using Boolean logic, not all Boolean logic statements can be written as a CART. For example, consider the logic statement (X_1 and X_2) or (X_1 and X_3). This says that the interaction between factors 1 and 2 or the interaction between factors 1 and 3 both lead to disease. As shown in Figure 2 below, this can be modeled exactly with logic regression but not with CART. In the CART model, the variable X_2 still must be assigned.

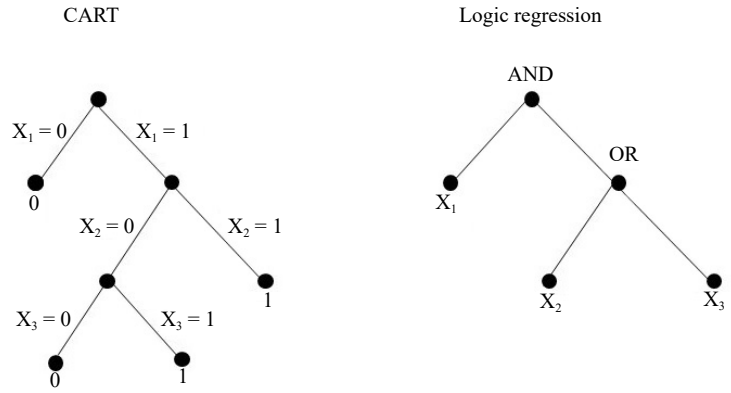

Figure 2. A comparison of a CART versus a logic regression tree. A CART predicts category 0 or 1, while the entire logic regression tree represents a prediction of category 1

While logic regression is more flexible than CART, it is not designed for the inclusion of continuous variables. C.Logic is an extension of logic regression that is especially effective in finding interactions of covariates. It uses joint dichotomization written about by Prince Nelson et al. combined with the Boolean logic mentioned above to model binary outcomes [5]. A future paper by Prince Nelson et al. will discuss the algorithm and theory of C.Logic. This paper shows through a simulation study, that C.Logic is superior to the Party package for CART in identifying interactions between continuous and binary variables in our specific framework

2. Methods

2.1 *Simulation study*

A simulation study was designed so that a true threshold existed in the data for a specific framework. For this study, we generated data so that observations were disease positive if there were interactions between X_1 and X_2 or X_3 and X_4 or X_5 and X_6. Note that variables X_4, X_5, and X_6 are continuous variables. There were also noise variables added. A true threshold, T, for the continuous variables was created so that there was a specific value for which the probability of disease increased. The prime implicant, L, was generated such that only specific interactions among the variables lead to an increased risk of disease.

A summary of the data generation is as follows:

a. Variables X_1, X_2, X_3 ~Bern(0,.3)
b. Variables X_4, X_5, X_6 ~Norm(0,1)
c. Set threshold, T, for variables X_4, X_5, X_6 such that $P(X \geq T = 0.524) = 0.3$
d. Define L as $L = \begin{cases} 1 & \text{if } (X_1 \wedge X_2) \vee (X_3 \wedge X_4) \vee (X_5 \wedge X_6) \\ 0 & \text{Otherwise} \end{cases}$
e. Let Y be a binary outcome such that $P(Y = 1) = P(L = 1)\, P(Y = 1 \mid L = 1) + P(L = 0)\, P(Y = 1 \mid L = 0)$ where $P(Y = 1 \mid L = 1) > P(Y = 1 \mid L = 0)$
f. Noise variables: $X_{7\text{-}17}$~Bern(0,.5) and $X_{18\text{-}20}$~Norm(0,1)
g. Strength of association between Y and L at three levels: Odds ratio = 2, 4, and 8
h. Sample sizes: 200, 300, 500, 1000, 1500, and 2000

2.2 *Algorithm*

The specifics of the C.Logic algorithm will be discussed in a future paper. In general, the algorithm first separates the candidate predictor variables into binary and continuous. Then, for each continuous variable, it selects a threshold using joint dichotomization as discussed in *An evaluation of common methods for dichotomization of continuous variables to discriminate disease status* [5, 6]. Next, it dichotomizes the continuous variables of the original data set

with the new thresholds. Once all the continuous variables have been dichotomized, it uses the Boolean logic of logic regression to select variables for the final model.

3. Results

For this simulation study, 500 repetitions were performed at sample sizes of 200, 300, 500, 1000, 1500, and 2000 with varying strengths of correlation between disease and the prime implicant of odds ratio = 2, 4, and 8. In order to determine how often each algorithm retrieved the "correct" combination of interactions, this "correct" answer was determined specifically by how we generated the data. For this data generation, the probability of disease only increased if there was an interaction between X_1 and X_2 or X_3 and X_4 or X_5 and X_6, also called our prime implicant.

Figure 3 shows that when the strength of association is low at odds ratio of 2, both algorithms have a difficult time choosing to add the exact correct interactions to the final model, though C.Logic is always slightly better and gets better as the sample size increases. By "exact", we mean that the interaction of interest was identified in the model without any noise variables. For example, $X_1 \wedge X_2$ only, not $X_1 \wedge X2 \wedge X_{16}$.

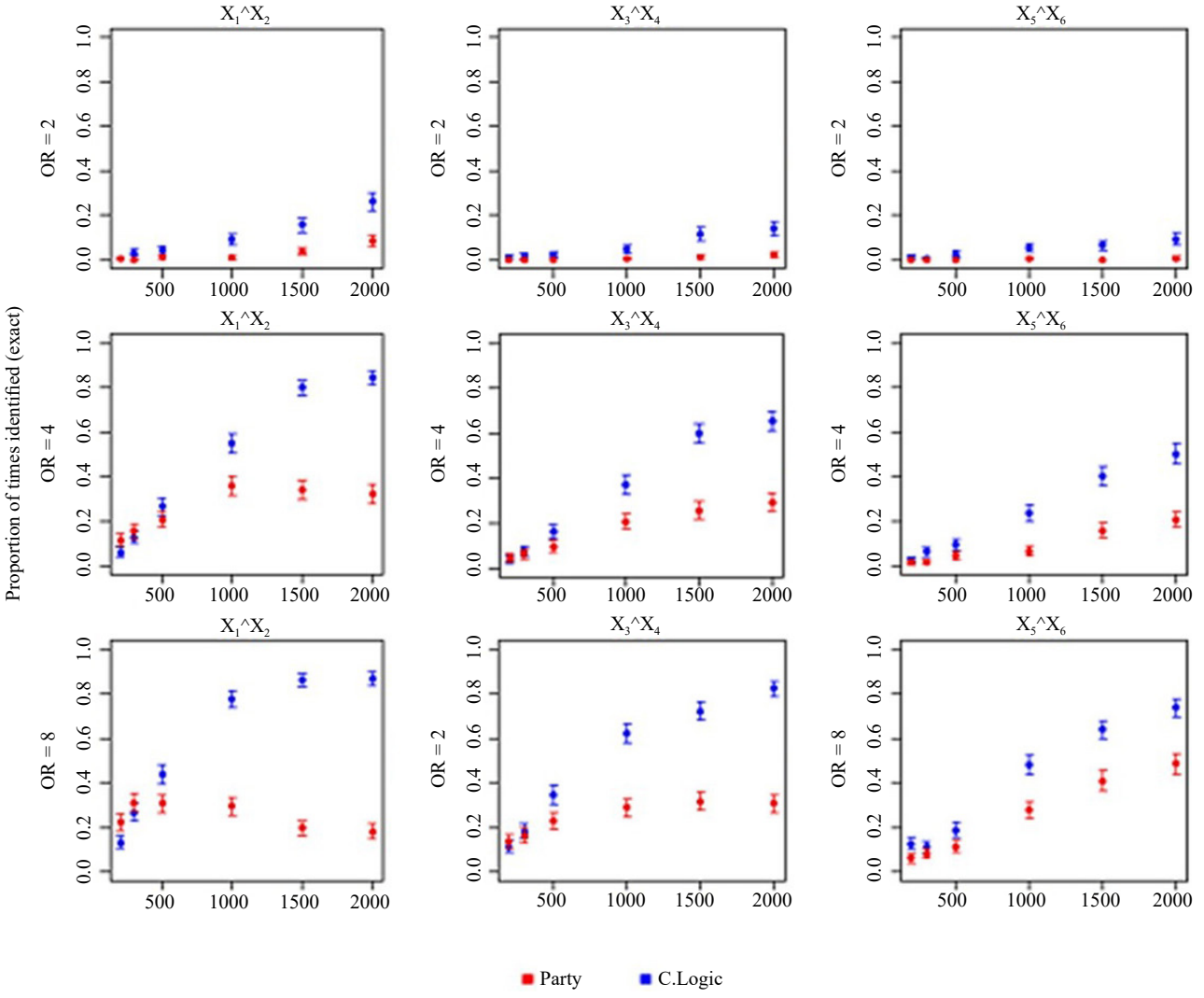

Figure 3. The results from a simulation study comparing the Party package to C.Logic. Each plot shows the sample size versus the proportion of times each algorithm correctly identified the interactions that increased risk of disease (OR: odds ratio)

At odds ratio of 4, both algorithms improve, but C.Logic performs significantly better than the Party package. This trend is more apparent as the sample size increases.

Finally, when the strength of association is increased to 8, C.Logic performs the best while the Party package is the same or perhaps only slightly better than it was at an odds ratio of 4. Once again, both packages are more accurate as the sample size increases, but C.Logic overwhelmingly outperforms the Party package at all levels.

4. Discussion

The purpose of the data generation framework was to simulate a situation where only the interactions of variables lead to an increased risk of disease not the main effects. This specific framework also takes into account the possibility that these interactions occur between binary variables, continuous variables and a combination of binary and continuous. This method of data generation has advantages and disadvantages. By design, a correct combination of interactions exists and thus we can compare how often each algorithm recovers the correct interactions. But because the framework is so specific, the results may not be generalizable to data sets where there are not interacting variables. As discussed in the introduction, however, it is more likely that the increased risk of disease is due to combinations of many factors not just the main effects [2-6]. Prince et al. also showed that the idea of simultaneous dichotomization, which is used in C.Logic, more accurately identifies variables that lead to increased risk of disease whether they are interacting with other variables or not [6]. Further, whether the interactions of variables occur between continuous variables, binary variables or both, the C.Logic algorithm is more effective in recovering them. This is seen in Figure 3 by how the C.Logic identifies the correct interactions more often than the Party package across all odds ratios and sample sizes and types of interactions.

5. Conclusion

Identifying interactions of variables that lead to increased risk of disease is important for many medical applications such as diagnosis and treatment. Traditional statistical methods such as logistic regression can accomplish identifying interactions in certain situations but will fail if the number of variables is too large or if there is an interaction without a main effect. Alternative tree-based methods are effective in the aforementioned scenarios but CART is biased toward the inclusion of continuous variables while logic regression is not designed for continuous variables at all. This paper compared two algorithms designed to address these issues. The Party package uses conditional inference to offset the continuous variable bias of CART while C.Logic uses simultaneous dichotomization in order to include continuous variables. Both Party and C.Logic improve at exactly identifying the correct interactions as sample size increases however, C.Logic identifies the interactions of interest more often than the Party package for every sample size and strength of association.

Acknowledgements

I would like to acknowledge my student Cierra Smith at the Citadel for helping to organize the information for this paper.

Conflict of interest

The authors declare no competing interests.

References

[1] Perkins NJ, Schisterman EF. The inconsistency of "optimal" cutpoints obtained using two criteria based on the receiver operating characteristic curve. *American Journal of Epidemiology*. 2006; 163(7): 670-675. https://doi.org/10.1093/aje/kwj063

[2] Lobo I. Epistasis: gene interaction and the phenotypic expression of complex diseases like Alzheimer's. *Nature Education*. 2008; 1(1): 180. https://www.nature.com/scitable/topicpage/epistasis-gene-interaction-and-the-phenotypic-expression-907/

[3] Manolio TA, Collins FS. Genes, environment, health, and disease: facing up to complexity. *Human Heredity*. 2007; 63: 63-66. https://doi.org/10.1159/000099178

[4] McKinney BA, Reif DM, Ritchie MD, Moore JH. Machine learning for detecting gene-gene interactions. *Applied Bioinformatics*. 2006; 5: 77-88. https://doi.org/10.2165/00822942-200605020-00002

[5] Prince Nelson SL, Ramakrishnan V, Nietert PJ, Kamen DL, Ramos PS, Wolf BJ. An evaluation of common methods for dichotomization of continuous variables to discriminate disease status. *Communications in Statistics – Theory and Methods*. 2017; 46(21): 10823-10834. https://doi.org/10.1080/03610926.2016.1248783

[6] Prince Nelson S, Ramakrishnan V, Nietert P, Kamen D, Ramos P, Wolf B. A comparison of joint dichotomization and single dichotomization of interacting variables to discriminate a disease outcome. *International Journal of Biostatistics*. 2022; 18(2): 613-625. https://doi.org/10.1515/ijb-2021-0071

[7] Loh WY, Shih YS. Split selection methods for classification trees. *Statistica Sinica*. 1997; 7(4): 815-840. https://www.jstor.org/stable/24306157

[8] Hothorn T, Hornik K, Zeileis A. Unbiased recursive partitioning: a conditional inference framework. *Journal of Computational and Graphical Statistics*. 2006; 15(3): 651-674. https://doi.org/10.1198/106186006X133933

[9] Ruczinski I, Kooperberg C, LeBlanc M. Logic regression. *Journal of Computational and Graphical Statistics*. 2003; 12(3): 475-511. https://doi.org/10.1198/1061860032238