DETERMINISTIC MODEL FOR CAUSALITY REASONING

Henry Su
Dr. Simon K. Poon
School of Information Technologies
THE UNIVERSITY OF SYDNEY

INTRODUCTION

At the heart of all complex causality analysis is the study of configurations of causal variables. One such approach, Boolean minimization, attempts to simplify logical expressions to extract key configurations. As we interpret these key configurations and understand how they capture the relationships between causal variables, inherent in the data, we begin to shift from causality analysis to causality reasoning.

Motivation

Previous to our work, the methodologies for Boolean minimization either had poor scalability, or were not deterministic. In particular, our work compares to a framework called CANAL, which utilises a selection strategies consisting of random choice, leading to non-deterministic results. Furthermore, its methodology largely stems from a statistical perspective, and neglected many of the concepts originating from the social sciences. We believed that through combining the concepts from both a statistical and a social science perceptive, an improved model for causal analysis could be developed - one that was deterministic, and more *accurate*.

Contribution

Our work improves upon the existing methodologies to deliver a deterministic and more accurate CAusality REasoning (CARE).

Its applications extends to many areas of research that utilises causality analysis or causality reasoning, such as health, finance, and social sciences. Whereas previous methodologies fall short in incorporating social concepts, CARE addresses these shortcomings, and furthermore, offers a deterministic and more accurate alterative.

METHODOLOGY

CARE consists of 4 stages: Social Scores, Coverage-Directed Search, Implicant Expansion, and lastly, Unate Covering. In comparison to its predecessors, CARE introduces a new pre-processing stage called Social Scores, which helps with selecting more accurate *literals*, compared to the previous strategy of using random choice. Also, whereas CANAL take an iterative approach to the causality analysis, CARE takes a heuristic-guided exhaustive approach.

Social Scores

The Social Scores is a pre-processing stage necessary for generating a set of scores for each causal variable. These scores are then later used as a heuristic to rank literals for the selection process in the following stages. Our algorithm for generating social scores involves generating a network graph from the original dataset, where each node represents a causal variable. From this, a score is calculated from a combination of the degree centrality and the betweenness centrality of a corresponding node.

literal: essentially a causal variable
implicant: a covering of one or more minterms of a Boolean function
minterm: a configuration of causal variables
onset: the set of minterms which lead to a desirable outcome
offset: the set of minterms which lead to an undesirable outcome

Coverage-Directed Search

The Coverage-Directed Search (CDS) finds a set of *implicants* which completely covers the onset.

1. The *onset* is initially uncovered.
2. Find an implicant which covers the largest proportion of the remaining uncovered onset and has the highest *social value*. Repeat this step until the entire onset has been covered.
   i. Recursively find and add literals, in order of highest *literal frequency* and *social score*. Stop when the term no longer intersects with the *offset*, at which stage an implicant is found. Previous strategies for selecting literals only used literal frequencies.
   ii. If there are multiple literal candidates with equal frequency and social score, all candidates are tried. Previously, a random choice was made, leading to non-deterministic results.

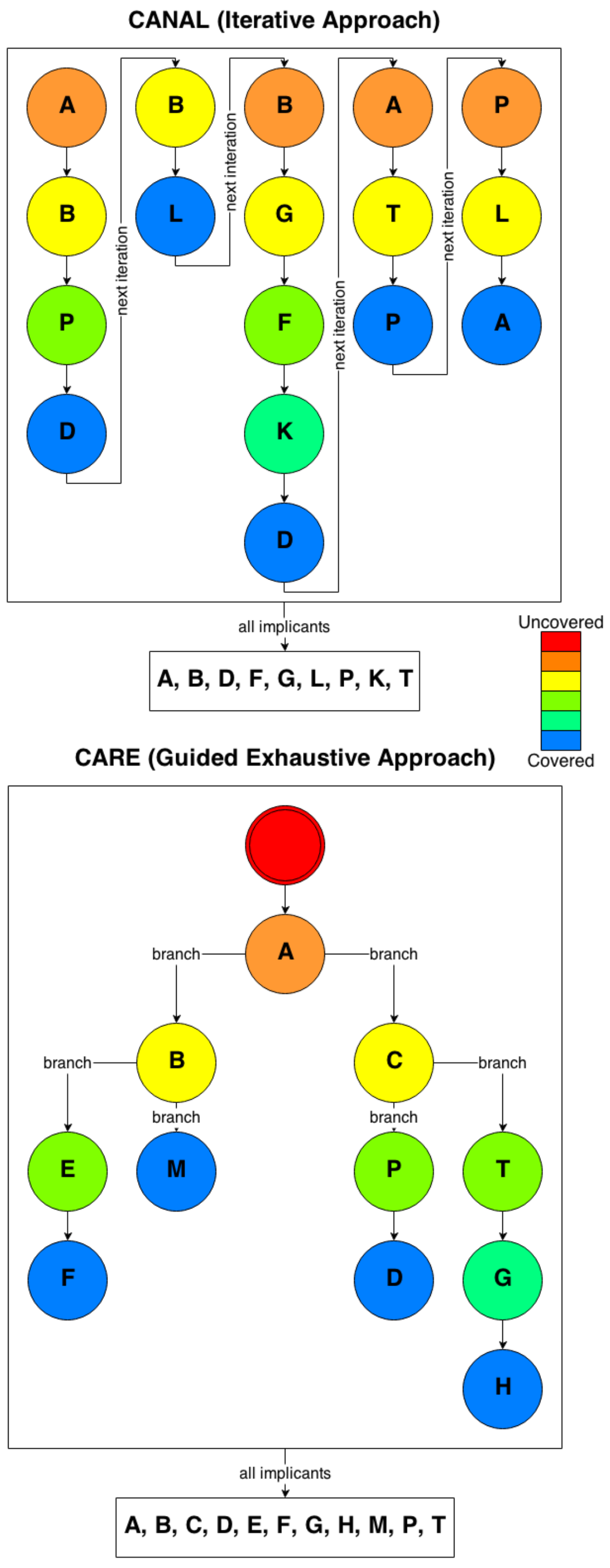

Diagram depicting processes in CANAL/ CARE

Implicant Expansion

The Implicant Expansion (IE) reduces each implicant, found in the CDS, to its minimal form, essentially creating a *prime implicant*. Our adaptation follows a similar exhaustive method used in the CDS. Apart from this, IE mostly remains faithful to the original method.

Unate Covering

Finally, the Unate Covering (UC) stage finds the minimal set of prime implicants which wholly covers the onset - that is, a set of *essential prime implicants*. This forms the final solution, and concludes the causality reasoning.

EVALUATION

Test Data

The hypothesis for the new methodology is that, by using social scores in the selection criteria, the literals which form each implicant more accurately captures the relationships embedded within the dataset. However, to test such a hypothesis, one must first know the relationships within the dataset.

1. Define a set of relationships for the causal variables. For comparison, generate a set of relationships for independent causal variables, highly interactive causal variables, and also a mixture.
2. Generate random configurations of the *minterms*, and label as onset of offset according the relationships. Repeat until a large enough pool of minterms have been generated.
3. Select a partial subset of random minterms from the pool, and use this to test the behaviours of CANAL and CARE.

Coverage

One aspect of the quality of the final solution is its *coverage*. Coverage is the proportion of the dataset covered by each EPI.

	CARE	*CANAL*
Mean	0.169557	0.162305
Variance	0.000142	0.000237
Observations	100	5
df	4	
t Stat	1.036838	
P(T<=t) one-tail	0.179188	
t Critical one-tail	2.131847	

t-test for onset coverage for the mixed dataset

These results demonstrate that CARE is capable of producing a solution with comparable coverage to that of CANAL. Also, note that CARE scored higher for the onset coverage, and lower for the offset coverage, suggesting that the new methodology produces more favourable results.

Accuracy

Another aspect of the quality of the final solution is its *accuracy*. Accuracy measures how well each EPI captures the original predefined relationships.

	CARE	*CANAL*
Mean	0.089209	0.062655
Variance	0.000165	3E-05
Observations	100	5
df	6	
t Stat	9.602524	
P(T<=t) one-tail	3.65E-05	
t Critical one-tail	1.94318	

t-test for accuracy for the interactive dataset

The t-test shows that the solution from CARE has a greater accuracy than that of CANAL, with 95% confidence. This reinforces the initial hypothesis that our methodology is more accurate at selecting literals.