LARGE SCALE CAUSAL COMPLEXITY ANALYSIS AND ITS APPLICATION IN TRADITIONAL CHINESE MEDICINE

Alan Su

Supervisor: Dr. Simon Poon, School of Information Technologies
FACULTY OF ENGINEERING & INFORMATION TECHNOLOGIES

ABSTRACT

The complex nature of prescriptions in TCM makes it extremely difficult to ascertain the true core herbs in prescriptions, that is, herbs that have a measurable effect on the treatment outcome, as opposed herbs that only serve a cosmetic effect. As a result, we need to perform causal complexity analysis in order to obtain the causal configurations.

Traditionally, analysis of causality has relied heavily on a correlational approach such as multivariate regression. However, it has been demonstrated by many researchers that such an approach cannot account for *conjunctural causation*, *equifinality* and *causal asymmetry*.

CONTRIBUTIONS

We introduce a novel approach for identifying the configurations which are of importance using the technique known as Boolean minimisation utilising an algorithm called BOOM. Unlike other Boolean minimization techniques, BOOM enables the analysis of datasets vastly in excess of what is capable to be processed by more traditional approaches.

Furthermore, we propose several extensions to the original algorithm to adapt it for use in causal complexity analysis and to enable the analysis of prescriptions not only accounting for whether if not a herb was present, but also the degree of a herb's presence by analyzing its prescribed dosage.

METHODOLOGY

The entire process of analyzing prescription data can be broken up into six discrete components:
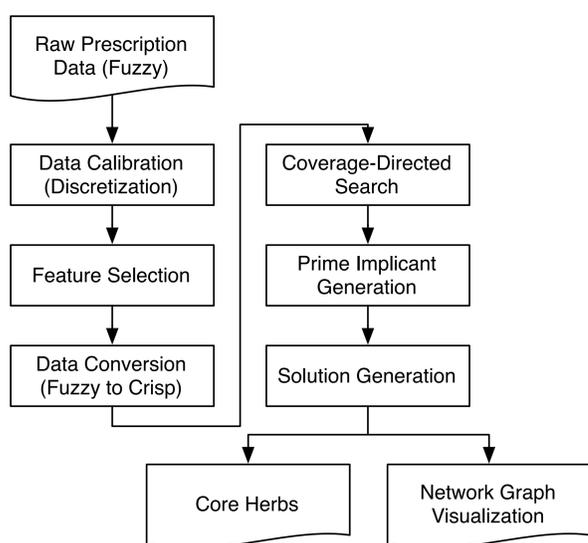

Figure 1: Data Analysis Process Workflow

Data Calibration

In order to assess the degree of membership of a herb in a prescription, calibration is performed using thresholds to discretize the data into a score that represents the degree in which a herb is present.

Suppose that we have the following five sets of prescription data shown in Fig. 2, whereby each contains up to five herbs and also their associated treatment outcome, the dosage levels are calibrated to conform to one of six discrete levels, 0.0, 0.2, 0.4, 0.6, 0.8 and 1.0, with 0.0 meaning that the herb was absent and 1.0 denoting that the herb was very heavily present in the prescription.

Herb A	Herb B	Herb C	Herb D	Herb E	Outcome
9.5g	0.0g	10.0g	0.0g	6.6g	0
0.0g	8.3g	10.5g	0.0g	10.1g	1
27.4g	4.0g	8.8g	0.0g	6.7g	0
11.8g	0.0g	6.0g	14.5g	6.3g	1
19.1g	5.9g	8.0g	2.1g	2.9g	1

Figure 2: Raw Prescription Data

Herb A	Herb B	Herb C	Herb D	Herb E	Outcome
0.2	0.0	0.8	0.0	0.6	0
0.0	0.8	0.8	0.0	0.8	1
0.8	0.2	0.6	0.0	0.6	0
0.4	0.0	0.4	0.8	0.6	1
0.6	0.4	0.6	0.2	0.2	1

Figure 3: Calibrated Prescription Data

Feature Selection

In order to perform causal complexity analysis with BOOM, the data must be encoded as either 0 for absent or 1 for present.

Our approach uses the method introduced by Charles Ragin to perform calibration. However, for datasets with large number of herbs, this becomes very computationally intensive due to the $O(e^n)$ complexity. As a result, the dimensionality of the data must be reduced before conversion.

Feature selection is performed using a slight variation of the BOOM algorithm to determine the herbs which are core to the analysis. These core herbs, which are a small subset of the original herb set, are kept to be used for the conversion process.

Data Conversion

Once the complexity of the prescription data is reduced, the calibrated data is then converted back into a crisp state.

Essentially, a new data matrix is generated whereby each row represents a prescription and each column represents a unique permutation of all the herbs, this is known as a *corner*. Prescriptions that contain n herbs will result in 2^n corners. From there, the membership score of the prescription against each of the 2^n corners is computed and the consistency of each corner is calculated using the following formula:

$$\text{Consistency}_{Corner} = \frac{\sum_{p \in \text{Prescriptions}} \min(\text{Outcome}_p, \text{Membership}_{C,p})}{\sum_{p \in \text{Prescriptions}} \text{Membership}_{C,p}}$$

If the consistency for a given corner exceeds a set threshold, the corner is assigned the outcome of 1. This subsequently results in a crisp dataset, with the herbs being the permutation and the outcome generated from the consistency calculation.

Coverage-Directed Search

The CD-Search is a component of BOOM which generates *implicants*, which are herbal configurations leading to a positive treatment outcome. Implicants are generated by adding herbs to candidates as long as the combination of herbs does not lead to a negative outcome.

An example of an implicant taken from the analysis of TCM prescriptions to treat insomnia is: ~太子参•~蜜甘草•~百合•~党参•~黄芩•~麦冬•~莲子心•石菖蒲•~淡豆豉•炒枳实•~薄荷•~浮小麦•~五味子•龙骨•~龙齿•~煅紫贝齿•~山药•~炒枳壳•~大枣•~山萸肉•~生甘草•~生姜•~炒白术•白芍•制远志•~知母•炒酸枣仁•牡蛎•~川芎

Prime Implicant Generation

Once the implicants have been generated, the algorithm attempts to generalize the implicants into *prime implicants* by removing herbs from implicants as long as the combination does not intersect with a prescription leading to a negative outcome.

By removing herbs from an implicant, it increases the coverage or applicability of that configuration to the entire dataset. An example of the corresponding prime implicant to the above implicant is: ~党参•~炒白术•白芍•炒酸枣仁•牡蛎

Solution Generation

The final step of the causal complexity analysis is to produce a minimal set of prime implicants which account for all positive treatment outcomes in the original prescription dataset. To accomplish this, the NP-Hard programming problem of *Unate Cover* is applied with a set of efficient heuristic functions to determine the solution. The final solution is a subset of all prime implicants.

RESULTS

1. ~浮小麦•炒白术•~炒酸枣仁•~柴胡•牡蛎
2. 淡豆豉•~大枣•生甘草
3. ~薄荷•煅紫贝齿•~炒白术•制远志•炒酸枣仁•~柴胡•~川芎
4. ~太子参•~黄连•石菖蒲•~薄荷•~法半夏
5. 白芍•炒酸枣仁•~柴胡•~川芎
6. ~茯苓•~黄芩•~麦冬•~生地黄•~山萸肉•~生甘草•~炒白术•~法半夏

Figure 4: Excerpt of Final Causal Configurations

By examining our approach using existing clinical TCM prescription data, we were able to compare the results with prior findings using other methods of analysis.

The results generated by our approach were relatively consistent with the results obtained through various means of analysis such that the core herbs were correctly identified and that the majority of herbal configurations generated were reasonable.

Our analysis tool also enables a visualisation of the relationships between the herbs building on the network graph analysis of the prescription data, an example is shown below:
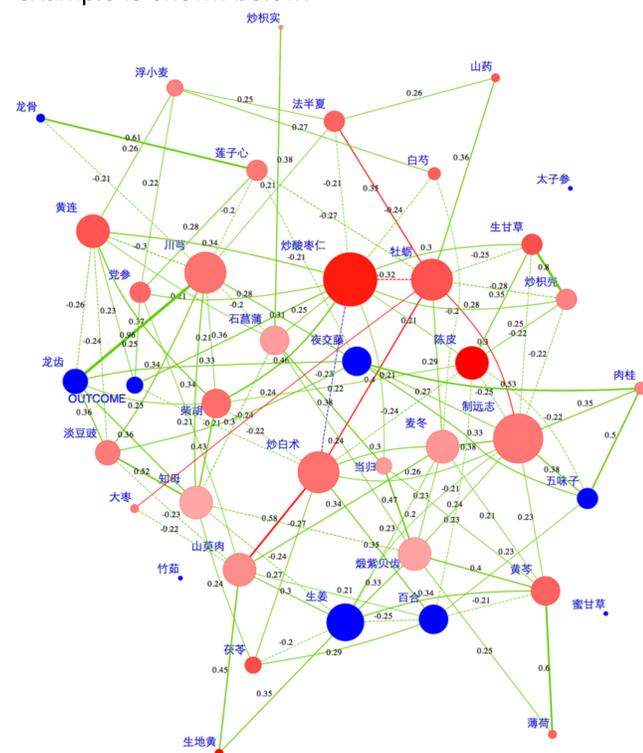
CONCLUSION

The goal of our research was to provide researchers with an analysis tool that enabled them to efficiently and accurately analyse TCM prescriptions to identify the inherent core herbs that are utilized across a vast number of prescriptions by different practitioners. This would enable medical researchers and practitioners to better understand the effective formulation of TCM prescriptions and to advance the field of Traditional Chinese Medicine.

Our approach demonstrates great potential to benefit the medical community as well as patients in search of better treatment.