Strategies to Build a Tree Summarizing Paths of Drug Use
Oscar Loureiro

Introduction

The objective of this document is to outline two different strategies to build a tree that summarizes the paths of drug use reported by the subjects interviewed for the Pathways project.

Definitions

Paths of drug use are defined as a sequence of stages. A stage is a moment in time (a year in the life of a subject) when the subject tries one or several new drugs.

For example, if a subject tried Alcohol and Exotic Marijuana at 15, Tranquilizers at 19, and Sniffed Heroin and Ecstasy at 21, then stage 1 would be trying Alcohol and Exotic Marijuana, stage 2 would be trying Tranquilizers, and stage 3 would be trying Sniffed Heroin and Ecstasy.

The number of stages varies greatly from individual to individual, with a minimum of 1 and a maximum of 7 (using 9 drug categories).

The nine drug categories are:

1 = A (alcohol)
2 = M(r/e) (marijuana re/ex)
3 = M&PCP (marijuana with PCP)
4 = M&H/C (marijuana w/h or c)
5 = H/C(s) (heroin or cocaine, s)
6 = H/C(in) (heroin or cocaine, in)
7 = T (tranquilizers)
8 = E (ecstasy)
9 = O (other)

But from this point forward only seven drug categories will be considered:

1 = A (alcohol)
2 = M(r/e) (marijuana re/ex)
3 = M&PCP (marijuana with PCP)
4 = M&H/C (marijuana w/h or c)
5 = H/C(s) (heroin or cocaine, s)
7 = T (tranquilizers)
8 = E (ecstasy)
**Data on Pathways of Drug Use**

As discussed in previous reports, the existence of ties (stages when more than one new drug is reported to have been tried) obscures the similarities among the paths followed by different subjects.

For example, consider three hypothetical subjects that followed exactly the same pattern of drug use: $A \Rightarrow M \ (r/e) \Rightarrow M\&PCP \Rightarrow T$. The first subject reports trying all 4 types of drugs when she was 14, the second reports trying $A$ at 12, the two types of marijuana at 14, and tranquilizers at 16, while the last one reports trying each type of drug at a difference age. In this case, if the ties were not resolved these three people will be reported as having followed three different paths, while resolving the ties will allow for the similarities in their paths to show.

The probabilistic or statistical approach used here to deal with ties is based on the idea of incorporating into the analysis the uncertainty that the ties represent.

For example, in the path $A\ and\ M(r/e) \ ? E$, we don’t know if the true path was $A\ ? M(r/e)\ ? E$ or $M(r/e)\ ? A \ ? E$, therefore both are equally likely to have happened to the subject in question. Then, the proposed approach would be to break the tie by counting $\frac{1}{2}$ subject having followed the path $A\ ? M(r/e)\ ? E$ and another $\frac{1}{2}$ subject having followed the path $M(r/e)\ ? A \ ? E$.

After this process of tie breaking, the data available is a series of paths followed by at least one of the subjects, each one with a frequency attached to it, the number of subjects (after breaking the ties) that followed that path.

Our goal is to summarize that long table of paths and frequencies (Appendix) into a tree that best represents the patterns followed by such paths.

**Two Strategies for Building Trees**

The first tree building strategy to be follow here is based on the approach used in Genetics to build phylogenetic trees. In that particular case, the original data is genomic information on different species and the goal is to build a tree where similar species appear in the same branch (e.g. gorilla and orangutan) while dissimilar species appear in different branches (e.g. gorilla and dog). The final result can be called a “similarity-based tree.”

The second tree building strategy to be followed here is based on the approach used in Computer Science to build data mining trees, e.g. from web browsing data. In that particular case no measure of similarity among subjects is considered (or even available). The goal is to find the tree of a certain size that leaves out the least amount of information contained in the recorded paths. The final result can be called an “information-based tree.”
Building a Similarity-Based Tree

Our goal is analytical: to establish how the path followed by a subject is related to some outcome of interest. To that end, we build a tree where subjects in the same branch tend to be similar regarding the outcome of interest and subjects in different branches tend to be dissimilar regarding the outcome of interest.

For example, let’s assume there are only 3 subjects, one with path ABC and outcome 14, another one with path BAC and outcome 12, and a third one with path CAB and outcome 4, where each letter represents a type of drug. Then, the appropriate similarity-based tree would be:

- A and B → C (values: 12, 14)
- C → A → B (values: 4)

A similarity-based tree is a graphical representation of how the different paths followed by the subjects are connected to the outcome of interest.

To build a similarity-based tree we start up with a tree that has a different branch for every path in the data and move towards a more compact tree by means of two operations: clustering and pruning.

We cluster two or more branches of a tree containing (among others) drug categories A and B when we fuse them into one containing the new drug category A and B. For example, if a tree looks like this:

- A → B → E → T
- B → A → E → T
- Other branches

and the subjects in the two upper branches have outcomes more similar between them than they are with respect to the other branches, then we can cluster those two upper branches to get the tree:

- A and B → E → T
- Other branches
We prune two or more branches of a tree containing (among others) drug categories A and B when we fuse them into one containing the new drug category A or B. For example, if a tree looks like this

\[ A \rightarrow E \rightarrow T \]
\[ \text{Other branches} \]

\[ \bigcirc \rightarrow B \rightarrow E \rightarrow T \]
\[ \text{Other branches} \]

and the subjects in the two upper branches have outcomes more similar between them than they are with respect to the other branches, then we can prune those two upper branches to get the tree:

\[ \text{A or B} \rightarrow E \rightarrow T \]
\[ \text{Other branches} \]

The usefulness of a similarity-based tree will be determined by having an appropriate outcome of interest and being able to measure it.

**Building an Information-Based Tree**

Our goal is purely descriptive: to build a tree of a certain size that contains as much of the information conveyed by the paths as possible.

What is the information conveyed by a path like \( A \Rightarrow M (r/e) \Rightarrow T \)?

The simplest unit of information conveyed by a path is a partial ordering between two categories, like “this subject tried A before trying T.” The path \( A \Rightarrow M (r/e) \Rightarrow T \) contains 3 of those partial orderings.

To build an information-based tree we will find the tree that, for a given size, contains as many of those units of information (partial orderings) as possible.

For example, supposed the information available is that 100 people followed the path ABCD, where each letter represents a drug category, 100 people followed the path ACBD, and 5 people followed the path DBCA. In other words, the information available is that:

<table>
<thead>
<tr>
<th>Ordering</th>
<th>Number People</th>
</tr>
</thead>
<tbody>
<tr>
<td>A before B</td>
<td>200</td>
</tr>
<tr>
<td>B before A</td>
<td>5</td>
</tr>
<tr>
<td>Condition</td>
<td>Frequency</td>
</tr>
<tr>
<td>-------------------</td>
<td>-----------</td>
</tr>
<tr>
<td>A before C</td>
<td>200</td>
</tr>
<tr>
<td>C before A</td>
<td>5</td>
</tr>
<tr>
<td>A before D</td>
<td>200</td>
</tr>
<tr>
<td>D before A</td>
<td>5</td>
</tr>
<tr>
<td>B before C</td>
<td>105</td>
</tr>
<tr>
<td>C before B</td>
<td>100</td>
</tr>
<tr>
<td>B before D</td>
<td>200</td>
</tr>
<tr>
<td>D before B</td>
<td>5</td>
</tr>
<tr>
<td>C before D</td>
<td>200</td>
</tr>
<tr>
<td>D before C</td>
<td>5</td>
</tr>
</tbody>
</table>

Then, the tree of size 4 (with 4 nodes) that contains more of this information than any other of that size is:

```
A       B
   \   /\    \
    C   \D
```

which contains 1000 of the 1230 orderings contained in the available paths.

**Appendix: Paths and frequencies for the new seven drug categories**

(....)