Event Pair Discriminativeness Ranking

Hao Li, Gregory Kramida, Andres Arellano, Michael Sorensen, Shobeir Fakhraei
University of Maryland,
College Park
hao.li.ict@gmail.com, gkramida@cs.umd.edu, andres.arellano@gmail.com, mdsoren@umd.edu,
shobeir@cs.umd.edu

1. PROBLEM STATEMENT AND FRAMEWORK STRUCTURE
The main problem that Multifinder addresses is: given any set of event sequences/records, \( S = s_1, s_2, ..., s_m \), composed of events from the set \( e_{all} = \{ e_a, e_b, e_1, e_2, \ldots e_n \} \), rank events based on which event is (i) most likely to cause outcome event \( e_b \) but not \( e_a \) (and vice versa) or (ii) most likely to result from \( e_a \) but not \( e_b \) and vice versa. It is not imperative for \( e_a \) and \( e_b \) to be mutually exclusive, i.e. a single record can include both.

We propose a simple framework for ranking such distributions of point events relative to two distinct outcome events. This framework combines several components indicating the amount of possible discriminativeness for distributions of each event type. Each one of these components relies on some measure that can serve as an indicator of causality for the two outcome events. We designate the value of a particular indicator \( j \) for event \( i \) causing outcome event \( x \) as

\[
I_j(i, x) = f_j(e_i, e_x)
\]

, where \( f_j \) is some black-box function indicating causality strength.

The basic measure of discriminativeness we derive for the \( i \)-th event using indicator \( j \) is thus:

\[
d_{ij} = |I_j(e_i, e_a) - I_j(e_i, e_a)|
\]

We are dealing with records and events from multiple domains, since the value of any indicator in determining the discriminativeness may vary from one domain to another or even from one event type to the other, we offer to weigh each discriminative measure with some user-defined factor \( \lambda_j \). The overall discriminativeness of the events is thus computed as:

\[
D_i = \sum_j \lambda_j d_{ij}
\]

2. CAUSALITY INDICATORS
We propose several causality indicators. It should be noted that any of the phenomena these indicators are based on are not protected against effects of confounding variables, i.e. if cooccurrence of the phenomenon with the outcome is caused by external event which influences both. We list the indicators below from most basic to most advanced.

2.1 Occurences Weighed by Time
Cheng et al., in their work on PairFinder \([?]\), suggest that the ratio of occurence count of any event \( e_i \) preceding an outcome event \( e_a \) (denoted here as \( n_{ix}^+ \)) and the occurrence count of \( e_i \) after the same outcome \( n_{ix}^- \) to the total count of occurrences is indicative of causality relationships between \( e_i \) and \( e_x \). Specifically, a high percentage of occurrences of event \( i \) before event \( a \) may be indicative of event \( i \) causing \( a \). Conversely, a high percentage of occurrences of event \( i \) after, but not before, event \( a \) may be indicative of event \( a \) causing event \( i \). The basic causality metric they present, using our notation is:

\[
I_{ratio}(i, x) = \frac{\text{sign}(n_{ix}^+ - n_{ix}^-)}{\frac{n_{ix}^- + n_{ix}^+}{\max(n_{ix}^-, n_{ix}^+)}} - 1
\]

, where the \( \text{sign}(y) \) yields 1 for any positive \( y \), -1 for any negative \( y \), and 0 when \( y \) is 0.

This function produces values in \([-1,1]\]. High negative values indicate a strong \( e_i \Rightarrow e_x \) relationship, whereas high positive ones indicate a strong \( e_x \Rightarrow e_i \) relationship.

However, when considering two possible outcomes \( a \) or \( b \), the occurence ratio metric is not usefull for event \( i \) in two cases: (i) when there are many records where both outcome events \( a \) and \( b \) occur in the same order relative to occurrences of \( i \), since the indicator values \([?]\) would be equivalent for cancel each other out during computation of discriminativeness in equation \([?]\), or (ii) when one or both of the compared outcome events occurs multiple times within each record, causing a conflict when occurences of \( i \) between any pair of outcomes are considered as both preceding and succeeding an outcome.

A reasonable solution to handle (i) is to convolve the
distribution of events $i$ for all records with some particular time distribution aligned relative to the outcome event, weighing each event occurrence with the value of the time distribution at time of occurrence. This follows the hypothesis that importance of some particular event occurring is often expected to have a particular importance distribution over time, $D(t)$. In many cases, for instance, events occurring more recently before or immediately after the outcome event may be considered more likely to have caused or to have been caused by it.

To put this in mathematical terms, let $t_{ikl}$ be the time of $k$-th occurrence for some event $i$ in a particular sequence/record. Then:

$$n_{ix}^- = \sum_l \sum_k \text{cnt}_{l,x}^-,$$

where

$$\text{cnt}_{l,x}^- = \begin{cases} D(t_{ikl}) & t_{ikl} < t_{xl} \\ 0 & \text{Otherwise} \end{cases}$$

Naturally, to compute $n_{ix}^+$ we need only to flip the inequality in (6). Note that this assumes a single occurrence of event $x$ per record.

Provided a non-uniform distribution of importance over time, the sum of occurrences weighed by time will then be tend to be different for the two outcomes even on the same record, so long as the two outcomes do not occur at exactly the same time. As shown in figure 1, the same weighing method may be used to alleviate some of the adverse effect of having multiple outcome occurrences per record:

Figure 1: $D(t)$ is shown in green around each of the outcome events $e_x$. Here, a normal distribution is used. This causes the $e_i$ occurrences between outcomes to weigh differently for the outcomes they precede and the outcomes they follow. Of course, this only works when proper distribution parameters, in this case $\sigma$, are chosen based on proximity of successive outcome occurrences.