Getting TAP patterns through sequence mining

Anmol Arora

August 29, 2016

1 Continuous TAPs

These are the general TAPs which start in some quantum and represent the following ordered sequence of states in consecutive fashion.
Given the data log, projection function π (σ , T, D_{raw}), quantization parameter τ and the state mapping function ψ, we can show that one can extract all continuous TAP sequences using sequence mining algorithm(like *PrefixSpan*).
Steps to get TAP sequences:

- Apply projection, quantization, aggregation and state mapping on the EET log.
- The assumption is that the transactions occur between two entities U and V. Consider the ordered transactions over quanta between any pair u, v : u $\in$ U, v $\in$ V as sequences.
  For example if u, v have edges labelled as:

  $$\tau_1 : S_2; \tau_2 : S_1; \tau_3 : S_3; \tau_4 : \text{(none)}; \tau_5 : \text{(none)}; \tau_6 : S_1$$

  Then the sequence contributed by this u-v pair is $S_2S_1S_3{*}{*}S_1$. '*'will be used when no edge is present in a particular quanta.
  Using this construction scheme, we get n*m sequences, each of length l, where n, m are the number of entities in both classes and l is the total number of quanta formed in the data log. These sequences form the sequence database D.
- Apply the sequence mining algorithm on the constructed D. Let the set of sequential patterns so obtained be SP. The claim is that set of continuous TAPs, hereafter called cTAP , form a subset of SP.

  $$\text{cTAP} \subseteq \text{SP}$$

  This follows from the definition itself. Any frequent substring(i.e a TAP) is also a frequent subsequence, hence the sequence mining process will find it.

- Next we need to find a method which will discard the sequences in SP $\setminus$ cTAP. Basically we need to check whether a given sequence $S \in SP$ occurs in at least α sequences in the database and starts in the same quantum in all of them.(Here α is the minimum support requirement).
- This can be done in following way: For each $Sq \in SP$, find the points of occurrence of substring Sq in S_{i} for $S_{\text{i}} \in D$ and mark them as in the figure below(The dots show the position where Sq starts). This can be easily done through a pattern matching algorithm such as KMP.

	τ_1	τ_2	τ_3	τ_4	τ_5	τ_6
S_1		•			•	
S_2	•		•			•
S_3	•		•		•	
S_4		•				•
S_5						
S_6	•		•			

If Sq is a valid TAP, then it must be true that for some quantum τ_{j}, it must start at τ_{j} in at least α sequences. Thus for each τ_{j}, we can simply count how many dots we encounter in that column(refer the figure above). If for some τ_{j}, the number of dots is more than α, then Sq is a valid TAP starting in quantum j. Else it is not a TAP and can be discarded.