How much variable is my event log?

Rowena Bernard

December 23, 2019

1 Introduction

A process variant can be defined as a group of executions of a process (i.e., a subset of cases of an event log) that share behavioral commonalities in terms of control-flow, performance and/or data attributes. It is important to study variability in event logs for the following reasons:

i. One of the several challenges in process mining is dealing with complex event logs which have diverse characteristics i.e variants, and various solutions which have been proffered are concerned with identifying and reducing variants [1].

ii. The decision to use a particular mining method for extracting models from logs in process discovery may be taken with regards to the variants or variability found in the log. [2]

iii. Process variability can be related to performance (e.g., throughput time of a process) or to other data attributes of events.

2 Defined variability metrics

i. Counting the number of variants :

The python function *compute_variant_variability()* takes as input an event log (path) and explores the event log using the PM4Py library

```
from pm4py.statistics.traces.log import case_statistics
```

and finally returns the number of distinct variants in log and a data-frame listing variants and their frequencies.

ii. Minimum Edit Distance:

Given an event log, (path) the function *compute_edit_distance_variability()* calculates the minimum edit distance between all sequence of traces in a event log.

Given two strings $a = INTENTION$ and $b = EXECUTION$, the edit distance of (a, b) is the series of edit operations (Insertion (I), Deletion(D), Substitution (S), that transforms a into b. One of the simplest sets of edit operations is that defined by Levenshtein in 1966:

```
I N T E * N T I O N
| | | | | | | | | |
* E X E C U T I O N
d s s   i s
```

- If each operation has cost of 1
  - Distance between these is 5
- If substitutions cost 2 (Levenshtein)
  - Distance between them is 8

The function *compute_edit_distance_variability()* first, extracts all traces from a given event log, and computes the edit distance between each trace and every other trace in the log. The total edit distance is averaged and result is given as output.

iii. Entropy: Entropy is a measure of the information required to represent an outcome of a stochastic variable and based on the exploratory study carried out by Christoffer et all (2019)[2], based on the exploratory study carried out, entropy measures may be used as a metrics of variability of an event log. But what is the random variable under consideration in the context of an event log? The function *compute_my_variability()* will compute **Trace Entropy**.

Trace entropy is defined by taking underlying random variable as ranging over entire traces, and using exact trace frequency as the probabilities.

Let $L = [T_1^{f1}, T_2^{f2}, ...T_n^{fn}]$ be a log with traces T_n, with individual frequencies, the trace entropy is computed as follows:

$$entropy(L) = -\sum_{T_i \in L} P_l(T_i) log P_l(T_i)$$

where $P(T_i) = \frac{f_i}{\sum f}$

3 Test Cases

We will run the complete python functions and show results for each log tested. The various logs used here can be found as labeled on the moodle page. The output of the functions:

```
logpath = 'L13.xes'
compute_variant_variability(logpath)
compute_edit_distance_variability(logpath)
compute_my_variability(logpath)
```

```
Number of variants in log:  4
Dataframe of variants and frequencies:
      variant  count
0    a,b,d,e     80
1    a,d,b,e     30
2  a,b,c,d,e     10
3        a,e      2

 Average Edit Distance between all variants
0.9483809781872375

Entropy (base2) and  Entropy (base10)
1.2899156456930654 0.38830330122988493
```

```
logpath = 'L14.xes'
compute_variant_variability(logpath)
compute_edit_distance_variability(logpath)
compute_my_variability(logpath)
```

```
Number of variants in log:  3
Dataframe of variants and frequencies:
    variant  count
0  b,a,c,d      1
1  a,c,b,d      1
2  a,b,c,d      1

 Average Edit Distance between all variants
2.0

Entropy (base2) and  Entropy (base10)
1.584962500721156 0.4771212547196623
```

```
logpath = 'L15.xes'
compute_variant_variability(logpath)
compute_edit_distance_variability(logpath)
compute_my_variability(logpath)
```

```
Number of variants in log:  6
Dataframe of variants and frequencies:
      variant  count
0  a,d,e,c,f      1
1  a,d,e,b,f      1
2  a,d,c,e,f      1
3  a,d,b,e,f      1
4  a,c,d,e,f      1
5  a,b,d,e,f      1

 Average Edit Distance between all variants:  1.8

Entropy (base2) and  Entropy (base10)
2.584962500721156 0.7781512503836435
```

```
logpath = 'L16.xes'
compute_variant_variability(logpath)
compute_edit_distance_variability(logpath)
compute_my_variability(logpath)
```

```
Number of variants in log:  4
Dataframe of variants and frequencies:
    variant  count
0  a,d,b,e      1
1  a,c,b,e      1
2  a,b,d,e      1
3  a,b,c,e      1

 Average Edit Distance between all variants:  1.6666666666666667

Entropy (base2) and  Entropy (base10)
2.0 0.6020599913279623
```

4 RESULTS FOR THE BPI CHALLENGE 2011

Results from the python functions on the BPICHALLENGE2011 event log below shows a variant size of 981, with a highest frequency count of any particular variant to be 41. The average edit distance between all variants in the BPICHALLENGE2011 is 195.8, and Trace entropy measures at 9.6.

Comparing this results to the above logs previously tested, which have large difference in result output, it can be seen that the different metrics used in the python functions: counting variants, edit distance and trace entropy can be useful methods for measuring the variability of a log.

```
1 logpath = 'BPIChallenge2011.xes.gz'
2 compute_variant_variability(logpath)
3 compute_edit_distance_variability(logpath)
4 compute_my_variability(logpath)
```

```
Number of variants in log:  981
Dataframe of variants and frequencies:
                                                 variant  count
0    vervolgconsult poliklinisch,administratief tar...     41
1    vervolgconsult poliklinisch,administratief tar...     17
2    1e consult poliklinisch,administratief tarief ...     16
3    vervolgconsult poliklinisch,administratief tar...     10
4    vervolgconsult poliklinisch,administratief tar...      8
..                                                 ...    ...
976  1e consult poliklinisch,1e consult poliklinisc...      1
977  1e consult poliklinisch,1e consult poliklinisc...      1
978  1e consult poliklinisch,1e consult poliklinisc...      1
979  1e consult poliklinisch,1e consult poliklinisc...      1
980  1e consult poliklinisch,1e consult poliklinisc...      1

[981 rows x 2 columns]

 Average Edit Distance between all variants:  195.88194492325937

Entropy (base2) and  Entropy (base10)
9.625190588260299 2.89747108104899
```

References

[1] W. M. P. v. d. A. Alfredo Bolt and M. de Leoni. Finding process variants in event logs. *from book On the Move to meaningful internet systems : OTM 2017 conferences : confederated international conferences: CoopIS, CTC, and ODBASE 2017, Rhodes, Greece, October 23-27, 2017, proceedings, part I (*, 308:45–52, Sep 2018.

[2] S. S. T. Back, Christoffer Olling; Debois. Towards an entropy-based analysis of log variability. *Business Process Management Workshops*, 308:53–70, Sep 2018.