Kiel University
Christian-Albrechts-Universität zu Kiel
Faculty of Engineering
DEPARTMENT OF COMPUTER SCIENCE

Process Analytics

Prof. Dr. Agnes Koschmider
Dominik Janssen

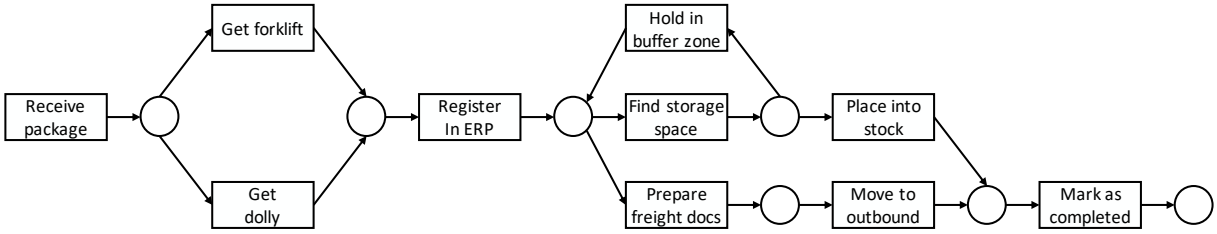
Advanced Process Mining

Summer term 2020

Exercise sheet 9

Event Log Clustering

Exercise 1: Trace Clustering



a) Cluster the process model above into

   i) 2 groups
   ii) 3 groups
   iii) 4 groups

b) What is the ideal number of clusters for this process model?

Solution

a) The process model can be clustered as follows:

2 groups	Heavy goods Regular goods
3 groups	Regular storing Storage full Ship directly
4 groups	Store heavy goods Ship heavy goods directly Store regular goods Ship regular goods directly

b) To determine an ideal number of clusters, the information given is not sufficient. The number of clusters depends very strongly on the scenario and the intention of the clustering.

Exercise 2: Distance Measures

a) Calculate the similarity between the following traces:

$$j = \langle A, B, C, D, E, F \rangle$$
$$k = \langle A, B, A, C, D, E, F, F \rangle$$

using the

i) Euclidean distance

ii) Hamming distance

b) In addition calculate the euclidean distance between the trace $m = \langle F, E, D, C, B, A \rangle$ and the trace j.
What is the drawback of this distance measure?

Solution

a) Based on *Trace Clustering in Process Mining* by Sung *et al.*:

$$n = 6$$
$$c_j = (1, 1, 1, 1, 1, 1)$$
$$c_k = (2, 1, 1, 1, 1, 2)$$
$$c_m = (1, 1, 1, 1, 1, 1)$$

i) Euclidean distance:

$$\begin{aligned} d_E(c_j, c_k) &= \sqrt{\sum_{l=1}^{n} |i_{jl} - i_{kl}|^2} \\ &= \sqrt{|1-2|^2 + |1-1|^2 + |1-1|^2 + |1-1|^2 + |1-1|^2 + |1-2|^2} \\ &= 1.4142 \end{aligned}$$

ii) Hamming distance:

$$d_H(c_j, c_k) = \frac{\sum_{l=1}^{n} \delta(i_{jl}, i_{kl})}{n} \qquad \delta(i_{jl}, i_{kl}) = \begin{cases} 0, & if(x > 0 \wedge y > 0) \vee (x = y = 0)) \\ 1, & \text{otherwise} \end{cases}$$
$$\begin{aligned} &= \frac{\delta(1,2) + \delta(1,1) + \delta(1,1) + \delta(1,1) + \delta(1,1) + \delta(1,2)}{6} \\ &= \frac{0+0+0+0+0+0}{6} \\ &= 0 \end{aligned}$$

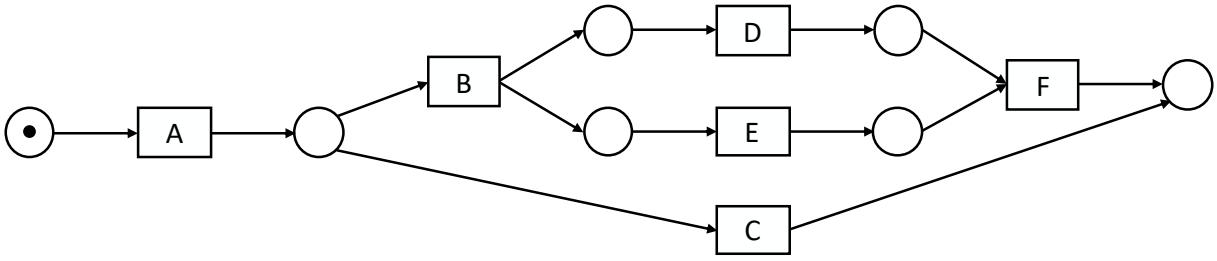
b) Euclidean distance:

$$\begin{aligned} d_E(c_j, c_m) &= \sqrt{|1-1|^2 + |1-1|^2 + |1-1|^2 + |1-1|^2 + |1-1|^2 + |1-1|^2} \\ &= 0 \end{aligned}$$

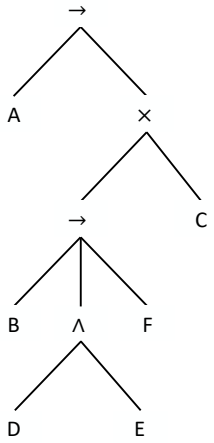
According to the Euclidean distance in conjunction with only counting the appearances of certain items, the cases j and m are similar. Yet m is in reversed order and not similar at all to j.

Exercise 3: Process Trees

Derive a process tree from the following Petri net:



Solution



Exercise 4: True or False

a) The k-Means clustering algorithm finds clusters of the size k.

b) The advantage of the Levenshtein distance over the Hamming distance is that it can be applied to vectors of different lengths.

Solution

a) **False.** The k stands for the number of clusters found by the k-means algorithm. The size of the clusters themselves are irrelevant.

b) **True.** The Levenshtein distance is calculated by counting the single-character edits necessary to get from one vector to the other. The Hamming distance, on the other hand, only allows substitution and can therefore only be applied to compare vectors of the same size. Calculating the Hamming distance can still be possible for traces of different lengths by converting the trace as it is done in exercise 2.