MIS Quarterly

BIG DATA & ANALYTICS IN NETWORKED BUSINESS

COMPREHENSIBLE PREDICTIVE MODELS FOR BUSINESS PROCESSES

Dominic Breuker
Hitfox Group GmbH, Rosa-Luxemburg-Str 2, 10178 Berlin, GERMANY {dominic.breuker@hitfoxgroup.com}

Martin Matzner
European Research Center for Information Systems (ERCIS), University of Muenster, Leonardo-Campus 3, 48149 Münster, GERMANY {martin.matzner@ercis.uni-muenster.de}

Patrick Delfmann
Institute for IS Research, University of Koblenz-Landau, Universitätsstraße 1, 56070 Koblenz, GERMANY {delfmann@uni-koblenz.de}

Jörg Becker
European Research Center for Information Systems (ERCIS), University of Muenster, Leonardo-Campus 3, 48149 Münster, GERMANY {joerg.becker@ercis.uni-muenster.de}

Appendix A

Synthetic Models and Their Characteristics

Table A1 summarizes the various characteristics of the synthetic models used in the experiments, including the number of event types, the size of the state space, whether a challenging construct is contained (loops, duplicates, nonlocal choice, and concurrency), and the entropy of the process defined by the model (estimated based on a sample of size 10,000). The original models may contain either duplicate tasks (two conceptually different transitions with the same label) or invisible tasks (transitions that have no label, as their firing is not recorded in the event log). We transformed all invisible transitions to duplicates such that, when there was an invisible task i in the original model, we added duplicates for all transitions t that, when fired, enable the invisible transition. These duplicates emulate the combined firing of t and i. Since we do not distinguish between duplicates and invisible tasks, we combined this category.

Table A1. Petri Net Models and Their Characteristics

Model	Events	States	Loops	Nonlocal Choice	Concurrency	Duplicates	Estimated Entropy
1	4	3	1	0	0	0	4.74
1Skip	4	6	1	0	1	1	5.97
2	4	4	1	0	0	0	1.97
2Optional	4	4	1	0	0	0	1.99
2Skip	4	5	1	0	0	1	2.02
a1	7	7	1	0	1	0	6.02
a10Skip	10	11	0	0	1	1	2.58
a12	12	13	0	0	1	0	2.25
a2	11	14	1	0	1	0	8.07
a5	5	6	1	0	1	0	2.35
a6nfc	5	7	0	1	1	0	1.50
a7	7	10	0	0	1	0	3.56
a8	8	8	0	0	1	0	1.92
betaSimplified	11	18	0	1	0	1	2.00
bn1	41	40	0	0	0	0	2.00
bn2	41	40	1	0	0	1	4.00
bn3	41	40	1	0	0	1	9.02
Choice	10	7	0	0	0	0	4.00
driversLicense	7	8	0	1	0	0	1.00
flightCar	6	8	0	0	1	1	1.92
herbstFig3p4	10	11	1	0	1	0	4.45
herbstFig5p19	3	6	0	0	1	1	2.51
herbstFig5p1AND	3	4	0	0	0	1	1.00
herbstFig5p1OR	6	8	0	0	1	1	1.00
herbstFig6p10	9	13	1	0	1	1	3.63
herbstFig6p18	5	5	1	0	0	1	6.77
herbstFig6p25	19	19	1	0	0	1	6.20
herbstFig6p31	7	7	0	0	0	1	2.00
herbstFig6p33	8	8	0	0	0	1	1.92
herbstFig6p34	10	15	1	0	1	1	6.44
herbstFig6p36	10	16	0	1	0	0	1.00
herbstFig6p37	14	51	0	0	1	0	9.25
herbstFig6p38	5	11	0	0	1	1	2.16
herbstFig6p39	5	11	0	0	1	1	3.42
herbstFig6p41	14	18	0	0	1	0	3.50
herbstFig6p42	12	20	0	0	1	1	3.95
herbstFig6p45	6	14	0	0	1	0	3.45
herbstFig6p9	5	7	0	0	0	1	2.00
parallel5	7	34	0	0	1	0	6.91

Detailed Results of the Experiments with Synthetic Data

Tables A2 and A3 document in detail the results of the experiments with synthetic data. As discussed in the section on experiment 2, we fitted a RegPFA to the training set (70%) of each of the event logs and measured the result's quality by computing the cross entropy with respect to the large test event log (10,000 process instances). The tables show the increase in cross entropy relative to the entropy of the actual entropy listed for each event log in Table A1. Therefore, the entries in Tables A2 and A3 represent the increase in entropy when the fitted model is used instead of the true model that generated the data. We report the performance with respect to each model's selection criterion and the "optimal" performance, that is, the performance that could have been achieved had the model selection delivered the best of all candidate models.

Table A2 lists results for the large event logs (700 process instances in the training set and 300 process instances in the validation set).

Table A2. Experiments on Synthetic Data with Large Event Logs

Model	Validation Set	AIC	$HIC_{0.05}$	Optimal
1	0.01	0.01	0.67	0.01
1Skip	1.02	1.88	2.71	1.00
2	0.00	0.00	0.00	0.00
2Optional	0.00	0.00	0.00	0.00
2Skip	0.00	0.10	0.00	0.00
a1	27.13	inf	27.01	20.76
a10skip	inf	34.45	28.23	27.04
a12	27.50	404.19	inf	19.70
a2	691.59	47.60	47.60	43.29
a5	0.01	0.00	0.00	0.00
a6nfc	12.51	42.62	5.87	3.76
a7	27.10	19.35	25.20	17.38
a8	inf	inf	18.30	9.27
betaSimplified	45.40	inf	43.11	36.63
bn1	0.00	26.27	0.00	0.00
bn2	inf	84.52	74.12	72.45
bn3	inf	145.69	111.74	107.00
Choice	24.83	inf	23.00	15.51
driversLicense	0.00	0.00	0.00	0.00
flightCar	138.49	26.68	26.68	21.91
hFig3p4	41.86	82.51	inf	38.48
hFig5p1AND	1.02	1.17	1.17	1.02
hFig5p1OR	0.00	0.00	0.00	0.00
hFig5p19	0.00	0.00	0.00	0.00
hFig6p10	40.68	inf	38.33	31.17
hFig6p18	inf	inf	15.11	14.39
hFig6p25	22.36	36.56	27.70	20.06
hFig6p31	421.27	421.27	16.50	11.12
hFig6p33	inf	inf	inf	13.76
hFig6p34	55.33	94.33	46.34	38.65
hFig6p36	1.00	1.00	1.00	1.00
hFig6p37	56.11	39.69	52.52	35.34
hFig6p38	98.71	362.66	inf	7.10
hFig6p39	22.67	25.92	20.51	17.58

hFig6p41	10.01	34.44	18.08	8.38
hFig6p42	34.54	33.59	33.27	25.41
hFig6p45	inf	16.21	15.40	12.06
hFig6p9	4.10	4.07	3.99	3.87
parallel5	0.10	3.21	0.25	0.10
# best choice	**16**	**11**	**24**	
# inf	**7**	**7**	**4**	**0**

Table A3 lists results for the small event logs (35 process instances in the training set and 15 process instances in the validation set).

Table A3. Experiments on Synthetic Data with Small Event Logs

Model	Validation Set	AIC	$HIC_{0.05}$	Optimal
1	inf	inf	10.84	3.57
1Skip	0.91	1.86	2.46	0.91
2	0.13	0.01	0.01	0.01
2Optional	0.32	6.07	0.01	0.01
2Skip	0.00	0.76	0.20	0.00
a1	25.16	29.00	24.08	13.59
a10skip	inf	27.55	25.42	22.50
a12	inf	19.33	7.14	5.18
a2	inf	51.33	36.88	31.09
a5	0.06	0.06	0.06	0.06
a6nfc	0.21	1.59	0.31	0.21
a7	25.80	9.09	1.74	1.14
a8	0.11	3.31	0.11	0.11
betaSimplified	inf	38.29	inf	28.02
bn1	inf	132.57	69.24	66.80
bn2	inf	198.52	73.48	69.74
bn3	inf	216.81	132.54	129.03
Choice	inf	15.77	inf	13.35
driversLicense	0.00	1.00	0.00	0.00
flightCar	0.07	7.89	0.07	0.07
hFig3p4	inf	36.64	inf	28.81
hFig5p1AND	17.66	12.30	16.37	11.89
hFig5p1OR	0.01	0.01	0.01	0.01
hFig5p19	0.00	5.12	0.00	0.00
hFig6p10	39.29	inf	34.31	31.51
hFig6p18	inf	107.95	14.93	11.98
hFig6p25	65.92	67.37	inf	59.24
hFig6p31	0.04	7.04	0.04	0.04
hFig6p33	inf	inf	22.87	18.66
hFig6p34	inf	52.81	inf	40.79
hFig6p36	1.03	7.03	1.04	1.03
hFig6p37	57.12	52.10	52.75	41.00
hFig6p38	18.96	80.53	22.53	6.20
hFig6p39	18.90	9.16	16.66	9.16
hFig6p41	0.23	11.75	0.24	0.23

hFig6p42	15.87	23.21	14.17	10.62
hFig6p45	5.94	inf	5.26	4.84
hFig6p9	inf	inf	15.65	11.56
parallel5	8.61	12.74	11.42	6.66
# best choice	**14**	**10**	**20**	
# inf	**14**	**5**	**5**	**0**

Table A4 shows the fitness and advanced behavioral appropriateness scores for all event logs used to evaluate the RegPFA Analyzer.

Table A4. Experiments on Process Discovery

Model	Fitness	Advanced Behavioral Appropriateness
|2|	1.00	0.69
|2|Optional	1.00	1.00
|2|Skip	1.00	0.59
a10skip	1.00	1.00
a12	1.00	1.00
a5	1.00	1.00
a6nfc	1.00	1.00
a7	1.00	1.00
a8	1.00	1.00
betaSimplified	1.00	0.65
Choice	1.00	1.00
driversLicense	1.00	1.00
driversLicensel	1.00	0.88
hFig3p4	1.00	0.74
hFig5p19	0.97	1.00
hFig6p18	1.00	0.81
hFig6p31	1.00	1.00
hFig6p36	1.00	0.80
hFig6p38	1.00	1.00
hFig6p41	1.00	1.00
Ø	**0.998**	**0.908**

Table A5 compares the experiment's results to the other algorithms' scores that de Weerdt et al. (2012) report.

Table A5. Comparison with Fitness and Advanced Behavioral Appropriateness Scores Reported in de Weerdt et al. (2012)

Algorithm	Fitness	Advanced Behavioral Appropriateness
ProbabilisticMiner	0.998	**0.908**
AGNES-Miner	0.995	0.813
α+	0.969	0.873
α++	0.984	0.879
DT Genetic Miner	0.996	0.778
Genetic Miner	0.998	0.737
HeuristicsMiner	0.973	0.809
ILP Miner	**1.000**	0.786

Appendix B

Description of the Baseline Predictors Used in Experiment 2

We applied n-gram models to business process event data in experiment 2. N-gram models, popular techniques for language modeling, distribute the event sequences of business processes by means of several conditional probability tables. For each sequence of up to n-1 events, a probability table is maintained that specifies the distribution over the next event. The distribution is modeled formally as follows:

$$P\left(X_0^{(c)},\ldots,X_{T_c}^{(c)}\right)=\prod_{i=0}^{T_c}P\left(X_i^{(c)}\middle|X_{i=1}^{(c)},\ldots,X_0^{(c)}\right)\approx\prod_{i=0}^{T_c}P\left(X_i^{(c)}\middle|X_{i-1}^{(c)},\ldots,X_{i-n+1}^{(c)}\right)$$

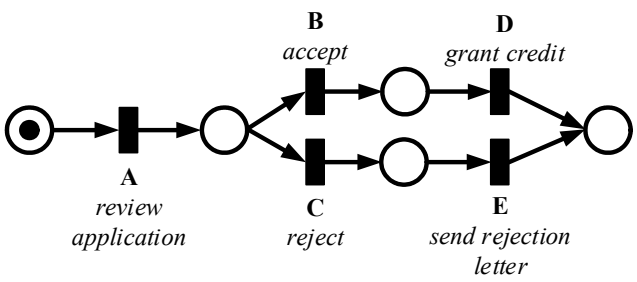
The conditional probability tables $P\left(X_i^{(c)}\middle|X_{i-1}^{(c)},\ldots,X_{i-n+1}^{(c)}\right)$ can be estimated by processing the event log to search for substrings that match the values of the $X_{i-1}^{(c)},\ldots,X_{i-n+1}^{(c)}$ variables and counting how often each event type follows the sequence. The counts allow probabilities to be estimated as relative frequencies.

As an example, Figure B1 shows the same business process and event log that Figure 1 shows, but it also contains conditional probability tables for a three-gram estimated from the five process instances in the event log. For instance, these tables predict that, after an event sequence *AB*, an event of type *D* will follow with probability 1.0, and after an event sequence *BD*, an event of type *kill* will follow with probability 1.0. Event *kill* is the artificial event that indicates process termination.

We show only a subset of all possible conditional probability tables. For instance, there is no table for event sequence *AD* because the tables are constructed from the relative frequencies with which certain types of events follow on the event sequence in the event log, and there is no occurrence of *AD* in the event log.

We maintain tables not only for event sequences of length $n-1$, but also for shorter event sequences. In the example in Figure B1, we maintain a table for the empty sequence (--) and for the sequence that contains only an event of type *A* (-A). The empty sequence is needed in order to model the probabilities of seeing a given type of event at the beginning of the process, while the sequence of only an event of type *A* (-A) is needed since *A* was observed in the event log with no event before it. In two out of five processes' instances, an event of type *B* follows after seeing only *A*. Three out of five instances proceed with a *C*, so the corresponding probabilities in the table are 0.4 and 0.6.

The *History* predictor, which we used in experiment 2 in addition to the n-gram, can be interpreted as a special type of n-gram. Since the *History* predictor is not limited in terms of the length of the event sequence it considers, it is an n-gram of unbounded length. Given a particular event log in which the longest process instance is of length $T_{\max} = \max_c T_c$, the *History* predictor is a $T_{\max}$-gram.



Event log

Process instance	Event sequence
C1	ABD
C2	ACE
C3	ABD
C4	ACE
C5	ACE

Sequence = (--)

Event type	Probability
A	1.0
B	0.0
C	0.0
D	0.0
E	0.0
kill	0.0

Sequence = (-A)

Event type	Probability
A	0.0
B	0.4
C	0.6
D	0.0
E	0.0
kill	0.0

Sequence = (AB)

Event type	Probability
A	0.0
B	0.0
C	0.0
D	1.0
E	0.0
kill	0.0

Sequence = (BD)

Event type	Probability
A	0.0
B	0.0
C	0.0
D	0.0
E	0.0
kill	1.0

Sequence = (AC)

Event type	Probability
A	0.0
B	0.0
C	0.0
D	0.0
E	1.0
kill	0.0

Sequence = (CE)

Event type	Probability
A	0.0
B	0.0
C	0.0
D	0.0
E	0.0
kill	1.0

Figure B1. Estimating a Three-Gram for an Exemplary Event Log

References

de Weerdt, J., de Backer, M., Vanthienen, J., and Baesens, B. 2012. "A Multi-Dimensional Quality Assessment of State-of-the-Art Process Discovery Algorithms Using Real-Life Event Logs," *Information Systems* (37:7), pp. 654-676.