Business Process Event Log Transformation into Bayesian Belief Network

Titas Savickas *titas.savickas@dok.vgtu.lt*
Vilnius Gediminas Technical University
Vilnius, Lithuania

Olegas Vasilecas *olegas.vasilecas@vgtu.lt*
Vilnius Gediminas Technical University
Vilnius, Lithuania

Abstract

Business process (BP) mining has been recognized in business intelligence and reverse engineering fields because of the capabilities it has to discover knowledge about the implementation and execution of BP for analysis and improvement. Existing business knowledge extraction solutions in process mining context requires repeating analysis of event logs for each business knowledge extraction task. The probabilistic modelling could allow improved performance of BP analysis. Bayesian belief networks are a probabilistic modelling tool and the paper presents their application in BP mining. The paper shows that existing process mining algorithms are not suited for this, since they allow for loops in the extracted BP model that do not really exist in the event log, and presents a custom solution for directed acyclic graph extraction. The paper presents results of a synthetic log transformation into Bayesian belief network showing possible application in business intelligence extraction and improved decision support capabilities.

Keywords: Process Mining, Bayesian Belief Networks, Business Process Analysis.

1. Introduction

Business processes (BP) are the most important aspect of every business as they define an activity or a set of activities that must be accomplished to achieve the organizational goals. Modelling and analysis of the processes is critical in identifying existing processes and understanding the contributions of new processes to the system [1]. An important facility in analysis of BP is the BP simulation. It is an aid to decision making, it helps in risk reduction and helps management at the strategic, tactical and operational levels [4].

Real processes encountered in real systems contain characteristics that make them complex by nature [4]:

- System uncertainty and stochastic nature is at the core of the business and has to be understood and evaluated;
- Dynamic behaviour – business processes tend to change over time;
- Feedback mechanisms – behaviour and decisions made at one point in the process impact others in complex or indirect ways.

While there are many existing solutions for the business process analysis, most of them require a lot of preparation to be capable of obtaining any meaningful knowledge. Existing simulation tools require manual creation of custom simulation models. Analysts have to prepare business process models and manually transform them into custom simulation models annotated with data required for simulation.

Process mining techniques are used to discover and analyze business processes in an automated way. Using all kinds of recorded process data, process mining techniques attempt to automatically discover the structure and properties of the business processes that can be visualized in business process models.

Existing process mining solutions extract information from event logs and, for each business intelligence question, they require to iterate through the log each time. Also, the existing methods often extract models that do not semantically conform to existing processes as they allow transitions between activities that do not exist in the existing business process.

Bayesian belief network based approach could solve these problems, where Business process model is extracted from an event log in the form of directed acyclic graph and annotated with data contained in event log to obtain a Bayesian belief network to facilitate BP analysis and decision making.

The paper is structured as follows: Section 2 provides a high-level view of the overall framework for the proposed approach with the emphasis on the parts reflecting the research presented in this paper. Section 3 formulates a background for the approach by revising process mining principles and theoretical foundation for constructing Bayesian belief networks. Then, Section 4 describes problems with existing process mining algorithms for Bayesian belief network extraction and presents directed acyclic graph extraction of the proposed approach. Section 5 defined Bayesian belief network over extracted directed acyclic graph and it's event log and Section 6 presents experimental results of an exemplary case. Section 7 provides some insights on the validity of the approach. The paper concludes with related works in Section 8 and conclusions with future research in Section 9.

2. Proposed Approach

This research focuses on transformation of event logs into directed acyclic graph and then into Bayesian belief networks. The algorithm is presented in Fig. 1.

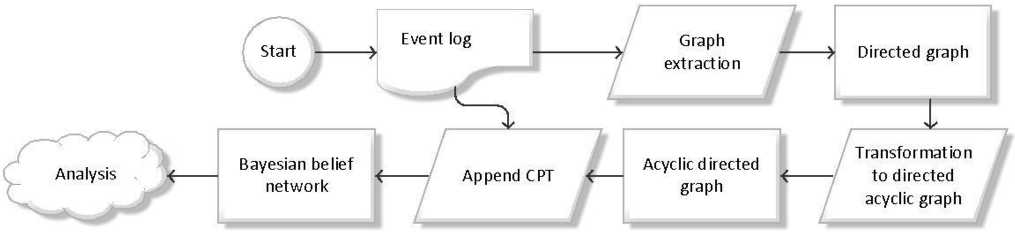

Fig. 1. Algorithm for event log transformation into belief network

The proposed approach starts with an event log (for details, see 3.1) extracted from existing information systems. The steps of the approach are:

1. Graph is extracted from the event logs. The graph contains all sequences of the traces in the event log. The graph is not really suitable for business intelligence extraction, because it might contain edges between nodes (transitions between activies) that are semantically incorrect (i.e. it allows for a possible loop in the process when there are no individual process instance traces allowing it. For details, see section 4).
2. The directed graph is transformed into directed acyclic graph to remove any loops (transitions for instance level loops are simply removed while model level loops are removed using an algorithm explained in section 4. Instance level loops are removed, because Bayesian network is directed acyclic graph. Model level loops are removed to make the model conform to the existing business process.
3. The data in the event logs is transormed to conditional probability tables (CPT) and appended to the directed acyclic graph.

The result is a Bayesian belief network ready for business process analysis.

3. Background

This section's purpose is to present background information on elements used in the approach – event logs and Bayesian networks.

3.1. Event Logs

Process mining focuses on extraction of knowledge from event logs commonly available in today's information systems. One of the purposes of business process mining is to discover business process models or to check conformance of/enhance existing models [7]. The process mining starts from the events stored in information systems (i.e. transaction logs, audit trails, etc.). Event logs used in process mining can be two kinds – MXML [14] and XES[15].

For this paper, van Dogen definition of event logs [12] is adapted. Two additional elements – M and δ – are introduced. The definition is as follows:

Definition 1. *An event log over a set of activities A and time domain TD is defined as* $L_{A,TD} = (E, C, M, \alpha, \gamma, \beta, \succ)$, *Where:*

- *E is a finite set of events*
- *C is a finite set of cases (process instances),*
- *M is a finite set of attributes,*
- $\delta: M \rightarrow E$ *is a function assigning each attribute to an event,*
- $\alpha: E \rightarrow A$ *is a function assigning each event to an activity,*
- $\gamma: E \rightarrow TD$ *is a function assigning each event to a timestamp,*
- $\beta: E \rightarrow C$ *is a surjective function assigning each event to a case,*
- $\succ \subseteq E \times E$ *is the succession relation, which imposes a total ordering on the events in* E

3.2. Bayesian Belief Networks

Since business processes are by nature complex and stochastic, it's useful to analyze those using probabilistic methods. Probabilistic methods, if used right, can greatly support decision making and answer questions with uncertainty.

One of such tools is Bayesian network, whose purpose is to model and reason with uncertain beliefs [2]. Bayesian networks can be defined as:

Definition 2. *A Bayesian network over variables X is a pair* (G, Θ), *where:*

- *G is a directed acyclic graph over variables X;*
- Θ *is a set of conditional probability tables (CPTs).*

Bayesian network inference can be used to answer questions important to business. Bayesian network inference can be related to business questions as follows:

- Probability of evidence $P(X|e)$. It Can be used to answer *„What's the chance to see a claimant aged 10-20 years old"*?
- Most probable explanation $MPE(\boldsymbol{e}) = \text{argmax}_{\boldsymbol{x}} \Pr(\boldsymbol{x}, \boldsymbol{e})$. It can be used to answer *„What is most probable explanation for declined claims"*?
- Maximum a Posteriori Hypothesis $MAP(\boldsymbol{e}, \boldsymbol{M}) = \text{argmax}_{\boldsymbol{m}} \Pr(\boldsymbol{m}, \boldsymbol{e})$. It can be used to answer "*What's the most probable outcome of a claim if the claimant is aged 23 years old and made the claim in the district of Vilnius?*";

4. Directed Acyclic Graph Extraction

The goal of the first step is to extract a directed acyclic graph from an event log. An acyclic graph is extracted, because Bayesian belief network is an acyclic graph. The graph is associated with the event log, but leaves out data attributes for the next step.

Directed acyclic graph in this context is defined as follows:

Definition 3. *Directed acyclic graph over event log L is defined as* $T_L = (N, D)$*, where:*

- $N = \{n \in E\}$ *is a set of nodes found in an event log,*
- $D = \{N \times N: n_i, n_j \in N, N \in c, c \in C\ n_i \succ n_j \& n_i \nprec n_j\}$ *is a set of edges connecting nodes, whose representative events follow each other but do not form a loop and exist in the same trace;*

Existing business process mining approaches are not suitable for Bayesian belief network creation. This is because they allow edges in the graph that would form a loop. The loops in those graphs appear because:

- The business process contains a loop and it is extracted as such from the event log;
- The business process contains parallel activities which are extracted as sequences from the event log. This extraction renders the model semantically incorrect and modelled activities become causally dependent, while in reality they are independent.

For example, there may be traces “ABCD”, “ACBD” and “ABBCD”. The first two traces do now allow for the loop to exist, but standard process mining tools would extract it in a way to allow a loop (Fig. 3a). The correct way would be to model them as independent activities (Fig. 3b). For the third trace, the process allows the loop and it would look like shown in Figure 3c.

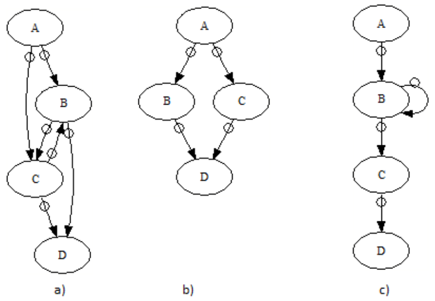

Fig. 2. Incorrect (a) and correct (b) graphs for traces “ABCD” and “ACBD”; graph of trace “ABBCD” (c)

In case of the presented approach, loops at the process instance level are simply ignored and left out of scope of this paper. This is done for two reasons – to simplify the task for initial research and because loops make up small part in business processes. The analysis of business process loops are left for future research. For the first case of loops, they need additional processing. The algorithm used in the approach to remove loops from a directed graph is shown in pseudo code below:

```
foreach (Trace t in EventLog)
  foreach(Event e in t)
    Graph.AppendNode(e,e.previousEvent,t)
Function AppendNode(Event e,Event previousEvent Trace t)
  if(TraceHasALoop(e,previousEvent,t))
    RemoveTransition(e,e.previousEvent) //instance loops are ignored
  if(ModelHasALoop(e,previousEvent))  //splits are added for model
loops
    possibleLoop=GetLoopPath(e,e.PreviousEvent)
    commonNode=GetNodeInPathToEprev(e,previousEvent,possibleLoop);
    commonNode.TransitionTo(e)
  else
    AddSimpleTransition(e,e.PreviousEvent);
```

The input for the algorithm is an event log with traces and it works on a directed graph. The output is a directed acyclic graph usable for Bayesian belief network generation. The generation is presented in the next section.

5. Bayesian Belief Network Generation

Directed acyclic graph is extracted from the event log, because the goal is to transform it into a Bayesian belief network and Bayesian belief network is a directed acyclic graph. The Bayesian Belief network itself is this context is defined as follows:

Definition 4. *Belief network over event log L is defined as* $B_L = (G, \theta, \varepsilon)$*, where:*

- G *is a directed acyclic graph* T_L*,*
- θ *is a finite set of conditional probability tables for nodes of graph G and their attributes M,*
- $\varepsilon\colon \theta \rightarrow N$ *is a surjective function assigning each conditional probability table to it's corresponding node;*

In the step 1, the extracted directed acyclic graph contains all the information necessary for Bayesian belief network, but the attributes are not aggregated into CPT. The second step creates a framework of the belief network - only the graph is extracted and the attributes are left out. The third step creates the CPTs and assigns them to the belief network. The final product of the process is a belief network ready for business process analysis.

6. Experimental Results

At the moment it is hard if not impossible to find publicly available event log with annotated data suited for business intelligence. Therefore, for experimental testing synthetic event logs were used. The logs were in XES format. The log is composed of 3437 traces consisting of up to 11 events. For the sake of simplicity, each event in the XES file was annotated with <string key="data" value="xxx">, where the value describes the data involved in the activity (i.e. age group of a client, call center location).

First, the directed acyclic graph using the presented approach and process model using PRoM framework were extracted to make sure the results are correct. The result of the proposed approach (Fig. 3 a) presented the most suitable graph for Bayesian belief network. Heuristics miner [16] (Fig. 3 b) contained loops both at trace level and model level. It was possible to tune the mining algorithm to remove the single activity loop, but the algorithm does not allow for complicated loop removal. Fuzzy miner [10] was also tested, but did not manage to achieve suitable level of detail because of looping or lack of connections. At the moment, there seems to be no mining algorithm suitable for the directed acyclic graph extraction, because they do not enforce model level loop checking.

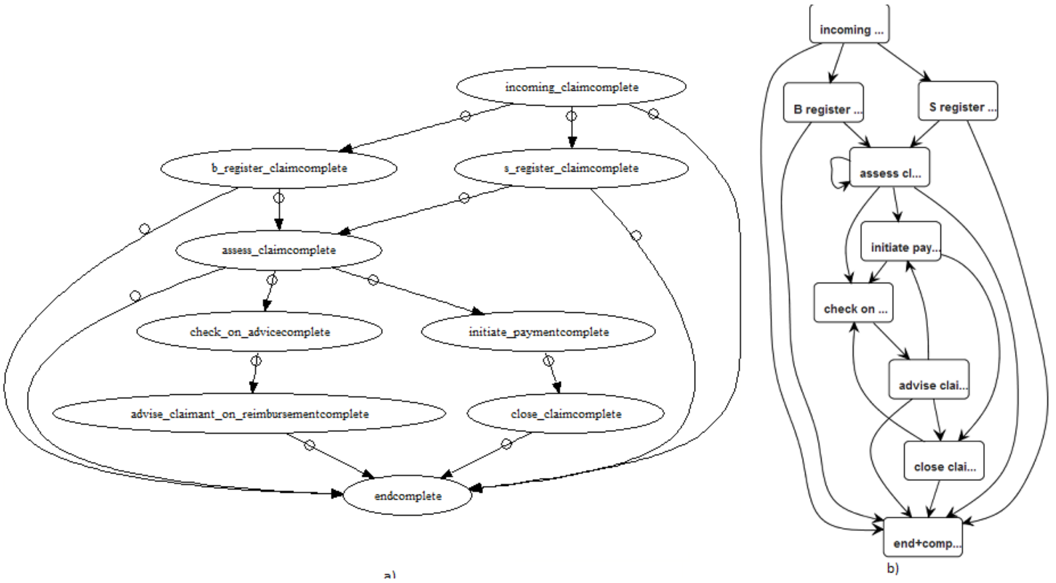

Fig. 3. Extracted graph from the synthetic log. a – proposed approach, b – heuristics miner.

The next step was to create a Bayesian belief network. To test the proposed approach, directed acyclic graph was transformed into factor graph and annotated with data existing in the event logs. For creation of the Bayesian network, Infer.NET framework [6] was used. The transformation went as follows:

1. Each node in extracted graph was directly transformed intto a variable node in the factor graph;
2. CPTs were created by analysing input edges of the nodes of the Bayesian network and extracted probabilities from process instance traces related to the particular nodes;

For readability purposes, only fragment of the resulting Bayesian network in factor graph form is presented. It is depicted in Fig. 4.

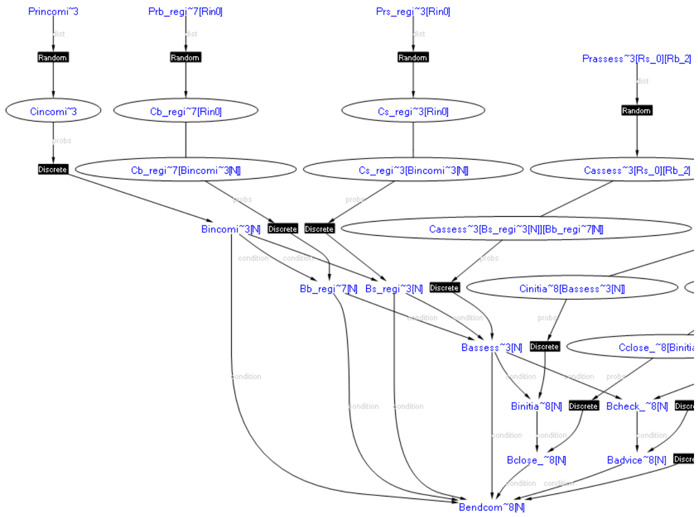

Fig. 4. Bayesian belief network in factor graph form

Finally, some possible questions about the process were answered. The questions about the process were chosen that would interest analysts and managers. The questions and results are presented in Table 1.

Table 1. Bayesian belief network inference results

Question	**Result**
Free form question: What's the chance of a claim to be declined? **Formal definition:** *P(endcomplete.data=’declined’)*	Answer: 0.36
Free form question: If the claim was registered in Klaipėda and it was declined, what's the most probable age group that claimed the insurance? **Formal definition:** *argmax P(incoming_claim.data|assessclaim.data= ‘Klaipėda‘ , endcomplete.data =’declined’)*	Answer: age 20-29
Free form question: If insurance was claimed by someone aged 50-59 years and the claim was registered in Vilnius, what are the possible outcomes? **Formal question:** *P(end | incoming_claim.dat =’50-59’,register_claim.Data =’Vilnius’)*	Declined: 0.43 Payed: 0.57

The questions that could be answered using the proposed approach could lead to the improvement of the business processes. For example, if the analysis of the business process shows that people aged 20-29 more often than not have unsuccessful insurance claim and people aged 50-59 more often than not have successful insurance claim, the designers of the business process or the managers could reallocate resources to make additional checks for younger people and less checks for the older people. The belief network could also be used to make temporal analysis and reduce the time it takes to complete the business process by making the necessary changes to the business process.

7. Validity

As it can be seen from the experimental results, the proposed approach allows automatic transformation of event logs into Bayesian belief network to support decision making. The approach shows promising results, but still needs formal validation. Some possible concerns for the experiment and the approach:

- The directed acyclic graph extraction algorithm worked for this scenario, but still needs additional checking and formal validation against existing techniques;
- Synthetic data was used and the process model is relatively simple. The approach needs to be checked against event logs that contain information about execution of dynamic processes or processes with a lot of loops;
- The proposed approach ignores any kind of loop at the process instance trace level. Analysts might want to know what causes the loops, but at the moment the proposed approach is not capable of incorporating information about loops in the extracted Bayesian Belief network.
- The proposed approach has not been validated against existing approaches therefore the conclusions of the performance of the approach are preliminary and shows linear scaling.

8. Related Work

The paper proposes custom graph extraction algorithm with a purpose of loop removal. There are many process discovery algorithms [13], but existing algorithms, i.e. HeuristicsMiner [16] or Fuzzy miner [10], extract graphs with loops that do not actually exist in the process rendering them unusable for generation of Bayesian belief networks.

Process mining has seen quite a few applications in business process analysis. It has been used for time prediction – in [12] authors use regression equations based on event logs to prepare model for prediction on when the process instance (case) will be finished; in [11] authors generate transition system from an even log which is used for time prediction of a case. Authors of [5] propose to simulate discovered models for use in decision support.

Process mining has also seen application in decision mining. In [8] authors attempt to extract rules for control flow point in the process model based on data in event logs. The rules are extracted using classification algorithms such as C.45. In [3], authors use alignment in business processes to extract data flow rules between activities.

Bayesian networks is not a new research area, but it’s application in process mining has not been widely researched. In [9], authors extract Bayesian network from an event log. The authors do not take into account the possible extracted structure, but rather attempt to extract it by analyzing dependency between events in the log. The approach presented in this paper uses custom network structure extraction to model dependency.

9. Conclusions and Future Work

This paper presented an approach for process analysis using process mining in combination with Bayesian belief networks. The paper highlighted problems with existing process mining algorithms and a custom solution for directed acyclic graph extraction which does not allow for any loops not existing in the event logs. For the scope of this paper, loops that do exist in the event log were ignored. It was shown how to transform the directed acyclic graph into Bayesian belief network to make decision support and analysis of the process simpler. The main conclusions are:

- The proposed directed acyclic graph extraction method allows extraction of process model that avoids loops at the model level. Further evaluation still needs to be done to check if it works for more complicated scenarios.
- The proposed Bayesian belief network transformation method shows promising results – the network can be created automatically from a BP event log; moreover, it allows for fast insight into process execution.

The approach presented in the paper is feasible; therefore, it can be further applied to more complicated case studies for verification. The future research involves validation using real life logs and improvement of Bayesian belief network transformation by using classification algorithms to account for continuous value ranges (i.e. age, currency, time values). The extracted Bayesian belief network shows only the data as seen by the execution of the process, it does not take into account “invisible data” (associated with activities but not shown in the event logs). It would be practical to combine the proposed approach with reverse engineering field to extract possible values, but not seen in the execution of the process. Finally, Bayesian belief network is a probabilistic model and it might have uses for simulation improvement.

References

1. Aytulun, S.K., Guneri, A.F.: Business process modelling with stochastic networks. Int. J. Prod. Res. 46(10), 2743–2764 (2008)
2. Darwiche, A.: Bayesian Networks. In: Van Harmelen, F., Lifschitz, V., Porter, B. (eds.) Handbook of Knowledge Representation, pp. 467-509. Elsevier, Amsterdam (2008)
3. De Leoni, M., Van der Aalst, W.M.P.: Data-aware process mining: discovering decisions in processes using alignments. Proceedings of the 28th Annual ACM Symposium on Applied Computing, pp. 1454–1461. ACM, Coimbra (2013)
4. Kellner, M.I., Madachy, R.J., Ra, D.M.: Software process simulation modeling : Why ? What ? How ? J. Syst. Software 46(2-3), 91-105 (1999)
5. Liu, Y., Zhang, H., Li, C., Jiao, R.J.: Workflow simulation for operational decision support using event graph through process mining. Decis. Support Syst. 52(3), 685-697 (2012)
6. Minka, T., Winn, J.M., Guiver, J.P., Knowles, D.A.: Infer.NET 2.5 (2012)
7. Rozinat, A., Mans, R.S., Song, M., Van der Aalst, W.M.P.: Discovering simulation models. Inf. Syst. 34 (3), 305–327 (2009)
8. Rozinat, A., Van der Aalst, W.M.P.: Decision Mining in Business Processes, http://www.processmining.org/_media/publications/beta_164.pdf
9. Sutrisnowati, R.A., Bae, H., Park, J., Ha, B.-H.: Learning Bayesian Network from Event Logs Using Mutual Information Test. Proceedings of the IEEE 6th International Conference on Service-Oriented Computing and Applications, pp. 356–360. IEEE, Koloa (2013)
10. Van Der Aalst, W.M.P., Rubin, V., Verbeek, H.M.W., Van Dongen, B.F., Kindler, E., Günther, C.W.: Process mining: a two-step approach to balance between underfitting and overfitting. Softw. Syst. Model. 9(1), 87–111 (2010)
11. Van der Aalst, W.M.P., Schonenberg, M.H., Song, M.: Time prediction based on process mining. Inf. Syst. 36(2), 450–475 (2011)
12. Van Dongen, B.F., Crooy, R.A., Van der Aalst, W.M.P.: Cycle Time Prediction: When Will This Case Finally Be Finished? In: Tari, Z., Meersman, R. (eds.) On the Move to Meaningful Internet Systems (OTM). LNCS, vol. 5331, pp. 319-336. Springer, Heidelberg (2008)
13. Van Dongen, B.F., De Medeiros, A.K.A., Wen, L.: Process mining: Overview and outlook of petri net discovery algorithms. In: van der Aalst, W.M.P., Jensen, K. (eds.) Transactions on Petri Nets and Other Models of Concurrency II. LNCS, vol. 5460, pp. 225–242. Springer, Heidelberg (2009)
14. Van Dongen, B.F.,Van der Aalst, W.M.P.: A Meta Model for Process Mining Data. Proceedings of the CAiSE Workshops (2005), http://tmpmining.win.tue.nl/_media/publications/dongen2005.pdf
15. Verbeek, H.M.W., Buijs, J.C.A.M., Van Dongen, B.F., Van der Aalst, W.M.P.: XES, XESame, and ProM 6. In: Proper, E., Soffer, P. (eds.) Information Systems Evolution. LNBIP, vol. 72, pp. 60-75. Springer, Heidelberg (2011)
16. Weijters, A.J.M.M., Van Der Aalst, W.M.P., Alves de Medeiros, A.K.: Process Mining with the HeuristicsMiner Algorithm. BETA Working Paper Series, WP 166, Eindhoven University of Technology, Eindhoven (2006) http://is.tm.tue.nl/staff/aweijters/WP166.pdf