Deep Learning Next-Activity Prediction With Cluster-Based Input Data

Felix Wolff

February 20, 2019
Version: My First Draft

Business Process Technology Group

Master's thesis

Deep Learning Next-Activity Prediction With Cluster-Based Input Data

Felix Wolff

1. Reviewer	Prof. Dr. Mathias Weske Business Process Technology Group Hasso Plattner Institut
2. Reviewer	Prof. Dr. ???????????? ???????????? ????????????
Supervisors	Dr. Luise Pufahl and Dr. Feng Cheng

February 20, 2019

Felix Wolff
Deep Learning Next-Activity Prediction With Cluster-Based Input Data
Master's thesis, February 20, 2019
Reviewers: Prof. Dr. Mathias Weske and Prof. Dr. ???????????
Supervisors: Dr. Luise Pufahl and Dr. Feng Cheng

Hasso Plattner Institut

Business Process Technology Group
Prof.-Dr.-Helmert-Straße 2-3
14482 and Potsdam

Abstract

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

Abstract (different language)

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

Acknowledgement

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

This is the second paragraph. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

Contents

1 Introduction

> *The most important contribution of management in the 20th century was to increase manual worker productivity fifty-fold. The most important contribution of management in the 21st century will be to increase knowledge worker productivity—hopefully by the same percentage.*
>
> — **Peter Drucker**
> (American management consultant and professor)

1.1 Research questions

In my thesis, I want to investigate the synergies of combining the aforementioned approach of Francescomarino et al. for learning data clustering with LSTM neural networks as per Evermann et al. Case data attributes shall be used during model training and prediction, as Polato et al. [Pol+14] and Schönig et al. [Sch+18] demonstrated their usefulness.

This would contribute to a field of research which is currently being explored and where LSTM networks have been applied successfully on prediction problems with long-term dependencies [ERF16; Tax+17; Sch+18; GS05].

Furthermore, I want to determine how historical case log data is prepared best for learning, as only Schönig et al. has written a small subsection on this [Sch+18]. If time permits, I also want to investigate the potential of ensembles within this context, as they can potentially enlighten the user about the reason for a prediction. With neural networks it is hard to comprehend the reasons behind a prediction. Other types of models deliver better comprehensibility.

Throughout the document I will strive to meet recently demanded machine learning paper quality criteria [LS18].

The performance of the combined approaches shall be evaluated against the data from the Business Process Intelligence Challenges (BPIC) 2011, 2012 and 2017 [;]. This allows for comparison with the results of Francescomarino et al., Evermann et al., Tax et al. and Schönig et al. [Fra+18; ERF16; Tax+17; Sch+18]. The next steps and an approximate timeframe are shown in the table below:

1.2 Thesis Structure

2 Background

2.1 Topical location?

Check what better term there is?

2.1.1 Predictive process monitoring:

Process science revolves around managing and optimizing structured procedures, while the broad area of data science covers data mining, algorithmic analysis and predictive analytics. Bridging the gap between the two fields is process mining [Aal16, p.18]. It covers the three steps of model discovery, conformance checking and model enhancement [Aal16].

These three steps are focused on offline data. If one would like to avoid a certain process outcome or e.g. an SLA violation, a guess at future developments requires resorting to online data analysis. At this step, techniques from the domain of predictive analytics can be employed[1].

Predictive analytics brings together a variety of statistical techniques like data mining, predictive modelling, and machine learning in order to make predictions about future events throught the use of historical data. In the domain of business processes, statistical or machine learning models are trained with historical process execution logs and *target* a specific piece of data that should be predicted. This application is called predictive process monitoring and allows answering questions such as *Given the current state of things, will I still meet my SLA?* or *Given the current case state, how long is this case still going to take?*. The answers to such questions can give case managers the opportunity to intervene if a case takes an unwanted course or might fail to meet KPI requirements.

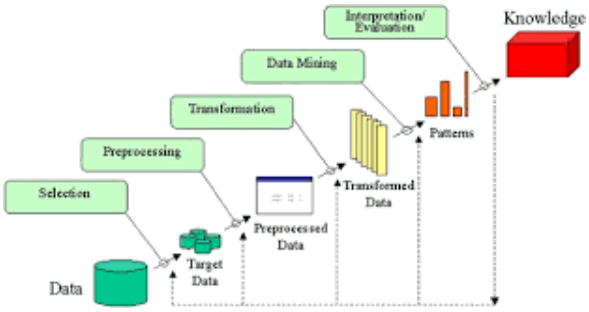

Figure 2.1: The process for *Knowledge Discovery in Databases*

2.2 Knowledge Discovery in Databases

Predictive analytics is a lot about model training, and the rough outline of necessary steps necessary to train a model is listed here:

1. Determine the *target* variable, it is the variable that is supposed to be predicted

2. Preprocess the dataset. This can mean introducing one-hot encodings, normalized values, but also basic data quality assurance such as null value elimination. Feature engineering can also happen at this step.

3. Partition the dataset into two parts. One part is set aside for model performance verification, as there the actual target variable value is known. This is commonly referred to as the *test set*, while the remainder is called the *training set*.

4. Train the model on the training set. Models are trained multiple times with different hyper-parameters to find the optimum configuration with respect to prediction accuracy on the test set. Hyper-parameters are model-specific values such as cutoff-thresholds that have impact on model performance.

2.3 Sequence prediction

2.3.1 Sequence-to-Sequence

2.3.2 Word-to-sequence

[1]More detail from Marlon Dumas on how these topics fit together: `https://www.youtube.com/watch?v=hMQolsRTOKO`

2.4 Sequence data inputs

2.4.1 Sliding Window

2.4.2 N-gram

2.4.3 Bag-Of-Words

2.4.4 Learned features, word2vec

Strictly piecewise
subsequences

2.5 Neural networks

2.5.1 RNN

2.5.2 LSTM memory

encoding decoding layers with long short-term memory An artificial neural network is an example for a machine learning model. It is made up of neurons, similar to its organic counterpart. The network is organized in three types of layers: a single input layer, one or more hidden layers and a single output layer. The neurons (being mathematical functions), pass their output on to those in the next layer via weighted connections. The weights on these connections are changed as the network is trained [Ros58]. Improving this forward-feeding network with backpropagation, i.e. learning from errors, made applications on pattern-detection successful[2]. Finally, enhancing the network with a way to remember sequences of events allows application on time-series data. The capacity as well as the durability of this memory are purposely limited as to avoid overfitting. As such, a long short-term memory inside a neural network functions similarly to our human one: we can remember a certain number of things for a short time, but we do forget some of them. The LSTM feature also equips the network with a remember and a forget capacity [HS97].

[2]Backpropagation can be attributed to many authors, as Schmidhuber blogs: `http://people.idsia.ch/~juergen/who-invented-backpropagation.html`

3 Related Work

> *A picture is worth a thousand words. An interface is worth a thousand pictures.*
>
> — **Ben Shneiderman**
> (Professor for Computer Science)

3.1 Sequence input formatting

3.2 Sequence prediction

3.3 Next-activity prediction

Hauder et al. mention numerous research challenges in the domain of ACM, among them an active support system for knowledge workers [HPM14]. The need for such a system is emphasized by Francescomarino et al. in their literature review, where it has been found that few prediction approaches target the next activity [Fra+18].

An example for how such a system might look like is given by Huber, who has developed a next-step recommendation system serving different case goals. The system is prototypically implemented into CoCaMa[1], a prototypical case management application. The system has been evaluated with 25 hand-made case logs.

Building upon each other are the works by Evermann et al. [ERF16] and Schönig et al. [Sch+18]. Evermann et al. have successfully demonstrated the good performance of long-short-term memory (LSTM) neural networks in predicting the next activity. Their approach did not take into account specific case data attributes however. How making use of this contextual information can improve the prediction accuracy even more, has been shown by Schönig et al. [Sch+18]. Furthermore Schönig et al. have explored data preparation methods for supporting the model during learning.

[1] CoCaMa is an abbreviation for a project called Collaborative Case Management, which appears to be retired: `http://archive.li/uZFnN`

Similarly, Polato et al. make use of environmental information in their work for improving the prediction of the remaining time of business process instances [Pol+14].

Metzger et al. predict run-time of a case by comparing and combining different prediction models into a model ensemble. Then, the members of the ensemble are selected based on their predictive performance measures. This allows taking into account costs of false predictions [Met+15].

Francescomarino et al. have performed clustering in the preprocessing phase of model training and prediction. Having clustered the training data, one model was created and trained for each cluster. For obtaining a prediction, the optimal cluster for a new data item is found from which the corresponding model is selected. This approach was evaluated on the accuracy of predicate fulfillment with two different clustering methods (k-means and DBSCAN) and two different prediction models (decision trees and random forests) [Fra+15]. A further evaluation criteria was *earliness*, i.e. at which point in time the correct result could be determined.

4 Contribution

> *Innovation distinguishes between a leader and a follower.*
>
> — **Steve Jobs**
> (CEO Apple Inc.)

4.1 sliding window with data

4.1.1 NULL Values

4.1.2 Staggering

4.2 sliding window with subsequence information

4.2.1 sliding window with data and prefixspan mined features

4.2.2 strictly piecewise features

5 Evaluation

> *Users do not care about what is inside the box, as long as the box does what they need done.*
>
> — **Jef Raskin**
> about Human Computer Interfaces

6 Conclusion

6.1 Verdict

6.2 Future Work

6.3 Acknowledgements

Bibliography

[] *BPI Challenge 2011*. https://data.4tu.nl/repository/uuid:d9769f3d-0ab0-4fb8-803b-0d1120ffcf54. Accessed: 2018-08-20 (cit. on p. 2).

[] *BPI Challenge 2012*. https://doi.org/10.4121/uuid:3926db30-f712-4394-aebc-75976070e91f. Accessed: 2018-08-15 (cit. on p. 2).

[] *BPI Challenge 2017*. https://data.4tu.nl/repository/uuid:7e326e7e-8b93-4701-8860-71213edf0fbe. Accessed: 2018-08-16 (cit. on p. 2).

[Aal16] Wil M. P. van der Aalst. *Process Mining: Data Science in Action*. 2nd ed. Heidelberg: Springer, 2016 (cit. on p. 3).

[ERF16] Joerg Evermann, Jana-Rebecca Rehse, and Peter Fettke. „A Deep Learning Approach for Predicting Process Behaviour at Runtime“. In: *Business Process Management Workshops*. 2016 (cit. on pp. 1, 2, 7).

[Fra+15] Chiara Di Francescomarino, Marlon Dumas, Fabrizio Maria Maggi, and Irene Teinemaa. „Clustering-Based Predictive Process Monitoring“. In: *CoRR* abs/1506.01428 (2015) (cit. on p. 8).

[Fra+18] Chiara Di Francescomarino, Chiara Ghidini, Fabrizio Maria Maggi, and Fredrik Milani. „Predictive Process Monitoring Methods: Which One Suits Me Best?“ In: *CoRR* abs/1804.02422 (2018). arXiv: 1804.02422 (cit. on pp. 2, 7).

[GS05] Alex Graves and Juergen Schmidhuber. „Framewise phoneme classification with bidirectional LSTM networks“. In: *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.* 4 (2005), 2047–2052 vol. 4 (cit. on p. 1).

[HPM14] Matheus Hauder, Simon Pigat, and Florian Matthes. „Research Challenges in Adaptive Case Management: A Literature Review“. In: *2014 IEEE 18th International Enterprise Distributed Object Computing Conference Workshops and Demonstrations* (2014), pp. 98–107 (cit. on p. 7).

[HS97] Sepp Hochreiter and Jürgen Schmidhuber. „Long Short-Term Memory“. In: *Neural Comput.* 9.8 (Nov. 1997), pp. 1735–1780 (cit. on p. 5).

[Les16] Stefan Lessmann. *Business Analytics & Data Science lecture*. Winter term 2015/2016.

[LS18] Z. C. Lipton and J. Steinhardt. „Troubling Trends in Machine Learning Scholarship“. In: *ArXiv e-prints* (July 2018). arXiv: 1807.03341 [stat.ML] (cit. on p. 1).

[Met+15] A. Metzger, P. Leitner, D. Ivanović, et al. „Comparing and Combining Predictive Business Process Monitoring Techniques“. In: *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 45.2 (Feb. 2015), pp. 276–290 (cit. on p. 8).

[Pol+14] Mirko Polato, Alessandro Sperduti, Andrea Burattin, and Massimiliano de Leoni. „Data-aware remaining time prediction of business process instances“. In: *2014 International Joint Conference on Neural Networks (IJCNN)* (2014), pp. 816–823 (cit. on pp. 1, 8).

[Ros58] F. Rosenblatt. „The Perceptron: A Probabilistic Model for Information Storage and Organization in The Brain“. In: *Psychological Review* (1958), pp. 65–386 (cit. on p. 5).

[Sch+18] Stefan Schönig, Richard Jasinski, Lars Ackermann, and Stefan Jablonski. „Deep Learning Process Prediction with Discrete and Continuous Data Features“. In: *Proceedings of the 13th International Conference on Evaluation of Novel Approaches to Software Engineering*. s.l., 2018 (cit. on pp. 1, 2, 7).

[Tax+17] Niek Tax, Ilya Verenich, Marcello La Rosa, and Marlon Dumas. „Predictive Business Process Monitoring with LSTM Neural Networks“. In: *CAiSE*. 2017 (cit. on pp. 1, 2).

[TPV09] Grigorios Tsoumakas, Ioannis Partalas, and Ioannis P. Vlahavas. „An Ensemble Pruning Primer“. In: *Applications of Supervised and Unsupervised Ensemble Methods*. 2009.

List of Figures

List of Tables

Colophon

This thesis was typeset with LaTeX 2_ε. It uses the *Clean Thesis* style developed by Ricardo Langner. The design of the *Clean Thesis* style is inspired by user guide documents from Apple Inc.

Download the *Clean Thesis* style at `http://cleanthesis.der-ric.de/`.

Declaration

You can put your declaration here, to declare that you have completed your work solely and only with the help of the references you mentioned.

Potsdam, February 20, 2019

Felix Wolff