Classifying business process execution traces with LSTM

Irene Teinemaa

December 30, 2015

1 Introduction

Process mining is a family of techniques to extract knowledge of business processes from event logs [1]. Two areas of process mining that make use of machine learning techinques are *deviance mining* and *predictive monitoring*. Deviance mining aims to explain the reasons why a process deviates from its normal or expected execution. Predictive monitoring aims to predict whether a running case will deviate or not, as early in the execution as possible.

Each execution (case) of a business process is a sequence of performed activities and, thus, sequence classification methods are applicable for deviance mining. In this report, I am using LSTM to classify between normal and deviant cases in a business process.

One drawback of this approach is that LSTM does not produce interpretable models, while an important aspect of deviance mining is the ability to *explain the reasons* why a process deviates. However, the given classification task can be thought of as a special case for predictive monitoring (the case has finished rather than running) and could be later extended to running cases.

2 Data

For evaluation, I use the Business Process Intelligence Challenge (BPIC) dataset from 2011. This event log records the treatment process of patients diagnosed with cancer in a Dutch hospital. The dataset contains 1140 traces and 622 unique activities. Following is a part of an example case with 12 activities: *outpatient follow-up consultation, administrative fee - the first pol, histological examination - biopsies nno, outpatient follow-up consultation, telephone consultation, assumption laboratory, assumption laboratory, unconjugated bilirubin, bilirubin - total, glucose, urea, hemoglobin photoelectric, ...*.

The labeling of the cases was chosen in accordance with [1], so that deviant cases are those where Diagnosis = "maligniteit cervix" and normal cases are all others. The ratio of deviant cases with this labeling is 0.196, so the dataset is rather imbalanced.

3 Evaluation

The evaluation was performed with 5-fold cross-validation. As the data set was imbalanced, I used oversampling on the training sets. To evaluate the performance of the classifiers, I measured accuracy and AUC.

The model was trained using the Keras library with Theano backend on the EEnet cluster. As basis, I used the example code for sequence classification with LSTM [1]. The following parameters were set: max_features = 622, maxlen = 200 (although some cases are longer), batch_size = 32, dropout = 0.2.

As baselines, I used decision tree and random forest, trained on a simple activity-occurrence matrix, where a value expresses how many times a given activity was performed in a given trace.

As can be seen from Table 1, the LSTM method did not improve over the baselines yet.

Table 1: Prediction results

Method	Accuracy	AUC
Decision tree	0.693	0.766
Random forest	**0.761**	**0.815**
LSTM	0.676	0.702

References

[1] Nguyen, Hoang, Marlon Dumas, Marcello La Rosa, Fabrizio Maria Maggi, and Suriadi Suriadi. *Mining business process deviance: a quest for accuracy.* In On the Move to Meaningful Internet Systems: OTM 2014 Conferences, pp. 436-445. Springer Berlin Heidelberg, 2014.

[1] `https://github.com/fchollet/keras/blob/master/examples/imdb_lstm.py`