Integrated Dimensionality Reduction and Sequence Prediction using LSTM

Poster · March 2018
DOI: 10.13140/RG.2.2.28577.30563

2 authors, including:

Lambert Schomaker
University of Groningen
220 PUBLICATIONS 4,282 CITATIONS

Some of the authors of this publication are also working on these related projects:

- Making Sense of Illustrated Handwriten Archives View project
- MPS - Medieval Paleographic Scale View project
Integrated Dimensionality Reduction and Sequence Prediction using LSTM

Emmanuel Okafor and Lambert Schomaker
Institute of Artificial Intelligence and Cognitive Engineering, University of Groningen, The Netherlands

Problem
- Most industrial or complex processes present temporal dependencies which stretch over a long time.
- The underlying patterns in these processes can be extremely non-linear.
- Use of linear predictive model (ARMA/ARIMA[1]) is not suitable.
- Hidden Markov Model[2] has prediction limitation when dealing with temporal dependencies that stretch over long durations.

Methods
- External Methods: Recurrent Neural Networks (RNN) [3-7]
- Proposed Method: Integrated Dimensionality-reduction LSTM

Results
- ID-LSTM Prediction on OHE codes during training and testing phases (left plot) and index predictions (right plot) over a duration of 10k time-counts.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Train</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID-LSTM-I-1-DIM-PCA-Codes</td>
<td>0.9760</td>
<td>0.9772</td>
</tr>
<tr>
<td>ID-LSTM-I-1-DIM-PCA-Codes</td>
<td>0.9424</td>
<td>0.9410</td>
</tr>
<tr>
<td>ID-LSTM-I-2-DIM-PCA-Codes</td>
<td>0.9463</td>
<td>0.9593</td>
</tr>
<tr>
<td>ID-LSTM-I-3-DIM-PCA-Codes</td>
<td>0.9424</td>
<td>0.9590</td>
</tr>
<tr>
<td>ID-LSTM-I-4-DIM-PCA-Codes</td>
<td>0.9139</td>
<td>0.9593</td>
</tr>
<tr>
<td>ID-LSTM-I-5-DIM-PCA-Codes</td>
<td>0.8729</td>
<td>0.9340</td>
</tr>
<tr>
<td>ID-LSTM-I-6-DIM-PCA-Codes</td>
<td>0.8715</td>
<td>0.9316</td>
</tr>
<tr>
<td>ID-LSTM-I-7-DIM-PCA-Codes</td>
<td>0.8715</td>
<td>0.9316</td>
</tr>
<tr>
<td>SL-LSTM-I-Raw-Codes</td>
<td>2.8580</td>
<td>0.0000</td>
</tr>
<tr>
<td>SL-LSTM-MSE-S1-Raw-Codes</td>
<td>2.8580</td>
<td>0.0000</td>
</tr>
<tr>
<td>SL-LSTM-MSE-S2-Raw-Codes</td>
<td>2.8580</td>
<td>0.0000</td>
</tr>
<tr>
<td>SL-LSTM-MSE-S3-Raw-Codes</td>
<td>2.8580</td>
<td>0.0000</td>
</tr>
<tr>
<td>SL-LSTM-MSE-S4-Raw-Codes</td>
<td>2.8580</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

NOTE
- A separate dimensionality reduction by PCA is not needed: the ID-LSTM uses 10 hidden dimensions in the bottleneck layer.
- One-hot-encoding is a must: do not try to predict arbitrary raw integer codes
- The left and right plots show the confusion matrix, that is the plot of the output predictions against their target values for both training and testing phases respectively for subset 9.

Conclusion
- We have transformed nominal codes to other vectorial representations with the objective of identifying correlated patterns using one hot encoding (OHE) and principal component analysis (PCA).
- Nominal integer codes are not sensible to use in the RNN.
- A separate dimensionality reduction by PCA is not needed: the ID-LSTM uses 10 hidden dimensions in the bottleneck layer.
- The ID-LSTM on OHE codes yield the best result on a small sample dataset.
- The use of ID-LSTM also obtains good results on reduced dimension PCA vector codes (20-DIM-PCA).
- The ID-LSTM obtained < 5% error on the predicted OHE codes in a realistically large dataset.

References

Future Directions
- We suggest that it may be possible to combine the proposed model with an early anomaly detection algorithm.
- To allow continuous prediction of physical problems in the machines generating the message logs.
- Optimization of LSTM-based feature dimensionality reduction in a realistically large dataset.