Integrated Dimensionality Reduction and Sequence Prediction using LSTM
Okafor, Emmanuel; Schomaker, Lambertus

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2018

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the “Taverne” license. More information can be found on the University of Groningen website: https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment.

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.
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 Handwritten Archives View project
- MPS - Medieval Paleographic Scale View project

All content following this page was uploaded by Lambert Schomaker on 10 April 2018.

The user has requested enhancement of the downloaded file.
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) is not suitable.
- Hidden Markov Model (HMM) has prediction limitation when dealing with temporal dependencies that stretch over long durations.

Objectives
- Use of external and a proposed integrated dimensionality reduction LSTM predictive systems for predicting message logs from industrial machines.
- Conversion of nominal codes (raw codes) to other vectorial paradigms to obtain better correlated patterns.

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.
- 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.

Table 1: Prediction accuracies for the different approaches for 10K Samples

<table>
<thead>
<tr>
<th>Methods</th>
<th>Train</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID-LSTM-I-OHE-Codes</td>
<td>0.9957</td>
<td>0.9920</td>
</tr>
<tr>
<td>ID-LSTM-I-2D-DIM-PCA-Codes</td>
<td>0.9763</td>
<td>0.9843</td>
</tr>
<tr>
<td>ID-LSTM-I-10-DIM-PCA-Codes</td>
<td>0.9316</td>
<td>0.9727</td>
</tr>
<tr>
<td>ID-LSTM-I-5-DIM-PCA-Codes</td>
<td>0.9139</td>
<td>0.9593</td>
</tr>
<tr>
<td>ID-LSTM-I-4-DIM-PCA-Codes</td>
<td>0.9424</td>
<td>0.9410</td>
</tr>
<tr>
<td>ID-LSTM-I-3-DIM-PCA-Codes</td>
<td>0.9463</td>
<td>0.9593</td>
</tr>
<tr>
<td>ID-LSTM-I-2-DIM-PCA-Codes</td>
<td>0.9424</td>
<td>0.9590</td>
</tr>
<tr>
<td>ID-LSTM-I-1-DIM-PCA-Codes</td>
<td>0.8729</td>
<td>0.9340</td>
</tr>
<tr>
<td>ID-LSTM-I-40-DIM-PCA-Codes</td>
<td>0.9767</td>
<td>0.9840</td>
</tr>
<tr>
<td>SL-GRU-MSE-SI-1-DIM-PCA-Codes</td>
<td>0.8715</td>
<td>0.9136</td>
</tr>
<tr>
<td>SL-GRU-MAE-SI-1-DIM-PCA-Codes</td>
<td>0.8715</td>
<td>0.9136</td>
</tr>
<tr>
<td>SL-LSTM-MSE-SI-1-DIM-PCA-Codes</td>
<td>0.8715</td>
<td>0.9136</td>
</tr>
<tr>
<td>SL-LSTM-MSE-SI-2-DIM-PCA-Codes</td>
<td>0.8594</td>
<td>0.9000</td>
</tr>
<tr>
<td>SL-LSTM-MSE-SI-3-DIM-PCA-Codes</td>
<td>0.8584</td>
<td>0.9000</td>
</tr>
<tr>
<td>SL-LSTM-MSE-SI-4-DIM-PCA-Codes</td>
<td>0.8584</td>
<td>0.9000</td>
</tr>
</tbody>
</table>

Table 2: Prediction accuracy of the ID-LSTM trained on OHE codes

<table>
<thead>
<tr>
<th>No of Subsets</th>
<th>Time counts</th>
<th>No of Index</th>
<th>No of Machine</th>
<th>Train</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subset 0</td>
<td>0-154M</td>
<td>948</td>
<td>20</td>
<td>0.9626</td>
<td>0.9751</td>
</tr>
<tr>
<td>Subset 1</td>
<td>154-309M</td>
<td>606</td>
<td>30</td>
<td>0.9979</td>
<td>0.9695</td>
</tr>
<tr>
<td>Subset 2</td>
<td>309-639M</td>
<td>535</td>
<td>36</td>
<td>0.9886</td>
<td>0.9624</td>
</tr>
<tr>
<td>Subset 3</td>
<td>639-1278M</td>
<td>619</td>
<td>48</td>
<td>0.9981</td>
<td>0.9021</td>
</tr>
<tr>
<td>Subset 4</td>
<td>1278-2556M</td>
<td>620</td>
<td>62</td>
<td>0.9837</td>
<td>0.9806</td>
</tr>
<tr>
<td>Subset 5</td>
<td>2556-5112M</td>
<td>675</td>
<td>109</td>
<td>0.9962</td>
<td>0.9347</td>
</tr>
<tr>
<td>Subset 6</td>
<td>5112-7680M</td>
<td>648</td>
<td>64</td>
<td>0.9205</td>
<td>0.9293</td>
</tr>
<tr>
<td>Subset 7</td>
<td>7680-15360M</td>
<td>679</td>
<td>95</td>
<td>0.9973</td>
<td>0.9576</td>
</tr>
<tr>
<td>Subset 8</td>
<td>15360-30720M</td>
<td>717</td>
<td>116</td>
<td>0.9943</td>
<td>0.9681</td>
</tr>
<tr>
<td>Subset 9</td>
<td>30720-61440M</td>
<td>624</td>
<td>263</td>
<td>0.9871</td>
<td>0.9268</td>
</tr>
<tr>
<td>Subset 10</td>
<td>61440-122880M</td>
<td>624</td>
<td>263</td>
<td>0.9844</td>
<td>0.9565</td>
</tr>
</tbody>
</table>

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.

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 reduceddimensional PCA vector codes (20-DIM-PCA).
- The ID-LSTM obtained < 5% error on the predicted OHE codes in a realistically large dataset.

References

This research was supported by the MANTIS Institute of Artificial Intelligence and Cognitive Engineering, University of Groningen, The Netherlands.