Long Short Term Memory Networks for Anomaly Detection in Time Series

1 Citation

Malhotra, Pankaj, et al. "Long short term memory networks for anomaly detection in time series." Proceedings. Presses universitaires de Louvain, 2015.

`https://www.elen.ucl.ac.be/Proceedings/esann/esannpdf/es2015-56.pdf`

2 Abstract

We stack LSTMs and make them predict the next few items in a time series. We then fit a Gaussian to the error terms and use the likelihood of the error terms as the outlier score.

3 Introduction

We have two models: LSTM Anomaly Detection (LSTM-AD) and RNN (with recurrent sigmoid units) Anomaly Detection (RNN-AD).

4 LSTM-AD: LSTM-based Anomaly Detection

Our time series is $X = \{\mathbf{x}^{(1)}, ..., \mathbf{x}^{(n)}\}$, where each $\mathbf{x}^{(t)} \in \mathcal{R}^m$. We select $d < m$ variables to predict and predict for the next l steps.

We split our set of normal sequences into train (s_N), validation-1 (v_{N1}), validation-2 (v_{N2}), and test (t_N) set. We split our set of anomalous sequences into validation (v_A) and test (t_A) sets.

We stack LSTMs on top of each other with fully connected feedforward connections between them. The input to the model is $\mathbf{x}^{(t)}$ and the output is a $\mathcal{R}^{d \times l}$ vector representing the next set of time series values. We train on s_N and use v_{N1} for early stopping.

We define error $\mathbf{e}^{(t)} = [e_{11}^{(t)}, ..., e_{1l}^{(t)}, ..., e_{d1}^{(t)}, ..., e_{dl}^{(t)}]$. Where $e_{ij}^{(t)}$ is the difference between $x_i^{(t)}$ and its value predicted at time $t - j$.

We fit a multivariate Gaussian $\mathcal{N} = \mathcal{N}(\mu, \mathbf{\Sigma})$ on the error terms from the validation and test sets and define $p^{(t)}$ as the likelihood of $\mathbf{e}^{(t)}$ under $\mathcal{N}$.

We classify a point as anomalous if $p^{(t)} < \tau$. We learn τ by maximizing the F_β score on v_{N2} and v_A. We pick $\beta = 0.1$.

5 Experiments

We use four datasets:

1. Electrocardiogram (ECG) data.
2. Time series measuring a valve on space shuttle.
3. Demand for power over time.
4. Multiple sensors on an engine time series.

Using $\beta < 1$ lets us prioritize precision over recall.

We find $p^{(t)}$ is low in anomalous regions, but there many be a few higher $p^{(t)}$ values here and there in the anomalous region.

The positive likelihood ratio (true positive rate divided by false positive rate) was very high (34).

You can see specific hidden units in the LSTM that only activate for an anomaly.

6 Discussion

LSTM-AD works better than RNN-AD.