Anomaly detection for time series using LSTM networks

Machine learning for time series mini-project

Mohammed Hssein [2] Soufiane Fafe [1]

[1]École Polytechnique & ENS Paris Saclay

[2]ENS Paris Saclay

April 27, 2022

Plan of the presentation

1. Introduction
2. Method
3. Data, experiments
4. Results
5. Conclusion
6. Bibliography

Introduction

- In data analysis, anomaly detection is generally understood to be the identification of rare items, events or observations which deviate significantly from the majority of the data and do not conform to a well defined notion of normal behavior.
- Anomaly detection has applications in many fields, such as system health monitoring, fraud detection, and intrusion detection. There are many types of anomalies.

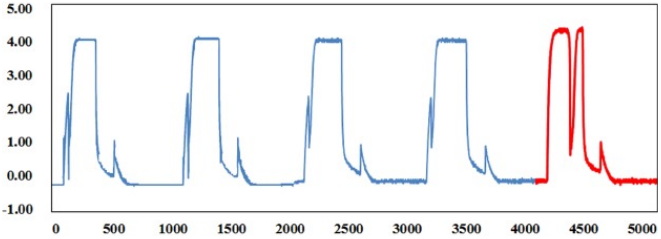

Figure: Anomaly in the red region

Method

The intuition about the method is simple:

- **Step 1** : Train the model to fit the normal time series behavior.
- **Step 2** : intuition Assume the model is well trained.
  - test data $\approx$ training data $\implies$ predictions close to real values
  - test data $\neq$ training data $\implies$ predictions far from real values
  - How to model this phenomenon in a probabilistic way ?
  - $\implies$ **Build** a probabilistic model for the errors conditioned on the model and data, by fitting the errors of the **normal test set** to a probability distribution, for example **Gaussian distribution** $\mathcal{N}(\mu, \Sigma)$
  - $\implies$ **Anomalous errors** should be in the tails of the distribution !
- **Step 3** : Anomaly detection incorporating the probability density learned in previous step
  - Observed error e is anomalous $\iff$ It is in the tail of a distribution , ie $\mathbb{P}(e|data, model) \leq \tau$

Model

We use a **Stacked LSTM** model:

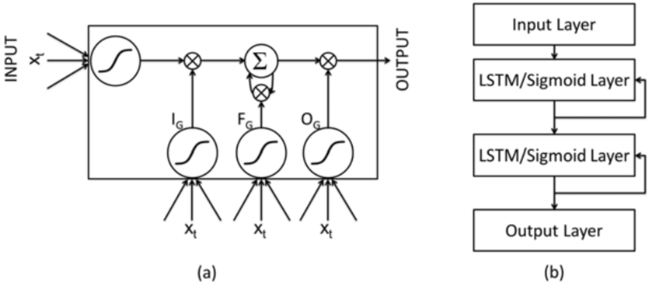

Figure: (a) LSTM network, (b) stacked architecture

Step 1: Training the Model

The model we use is a **Stacked LSTM** network:

- We use two LSTMs sequentially
- Train to predict l steps in the future using h steps from the past ($h = 5$ or 10, and $l = 3$ or 1).
- Hyper-parameters:
  - Batch-size: 64
  - Adam optimizer (learning rate 10^{-3})
  - Loss: Mean squared error
- Evaluating the model performance : Regression problem
  - Mean squared error
  - R2 score
- Training sets:
  - S_N v_{N_1} training and validation sets (with no anomalies)
  - t_N test set (with no anomalies)

Step 2 : Error modeling

- The model is trained now !
- **Build** probability distribution $\mathcal{N}(\mu, \Sigma)$ to **model** the behavior of normal data.
  - Make predictions on the normal test data t_N and compute error vectors $\mathbf{e} := \{\mathbf{e}_1, ..., \mathbf{e}_n\}$.
  - Fit $\mathbf{e}$ to $\mathcal{N}(\mu, \Sigma) \implies$ **Maximum Likelihood Estimation**
  - We have

$$\hat{\mu} := \frac{1}{n}\sum_{i=1}^{n} \mathbf{e}_i \ , \quad \hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} (\hat{\mu} - \mathbf{e}_i)(\hat{\mu} - \mathbf{e}_i)^T$$

  - This distribution models the normal behavior of the errors.
  - $\mathbb{P}(e|data, model) \leq \tau \implies$ anomaly (and vice-versa !)
- **Problem** : How to define τ ?
  - The choice of τ impacts the performance.
  - Incorporate performance metrics as **Recall** and **Precision**.
- **Solution** : τ is chosen to maximize the F_β score for small β ($\beta = 0.1$).

Step 3 : Anomaly detection

Now everything is ready ! We should detect anomalies in the anomalous data t_A. Note that we have:

$$\mathbb{P}(e|data, model) \leq \tau \iff -\log \mathbb{P}(e|data, model) \leq \log\left(\frac{1}{\tau}\right)$$
$$\iff \frac{1}{2}a(e) \geq 2\log\left(\frac{1}{\tau}\right) - 2C$$

where we have

$$a(x) = (\hat{\mu} - x)^T \hat{\Sigma}(\hat{\mu} - x)$$

is called the **Mahalanobis distance**.

Recap :

$$\tfrac{1}{2}a(e) \geq 2\log\left(\tfrac{1}{\tau}\right) - 2C \iff \text{e is anomalous.}$$

Datasets

We used 3 datasets in our experiments:

- Synthetic dataset : A periodic dataset we obtained by summing 3 sin functions with different frequencies (50, 100, 300 respectively). We engineered anomalous regions by changing the local mean and variance.
- ECG dataset :Electro-Cardiogram dataset. The original contains one anomalous interval. We studied the effect of adding more anomalie similar to the original one
- Respiration dataset : Dataset of normal and anomalous respiration of a human.

Results : ECG / Respiration datasets

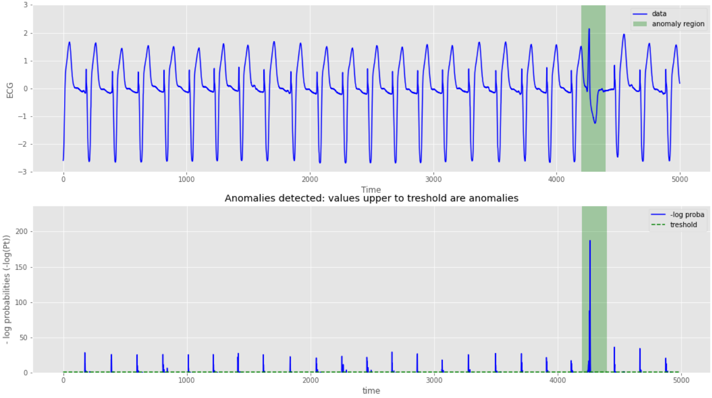

Figure: ECG dataset with only one anomaly interval

Results : ECG / Respiration datasets

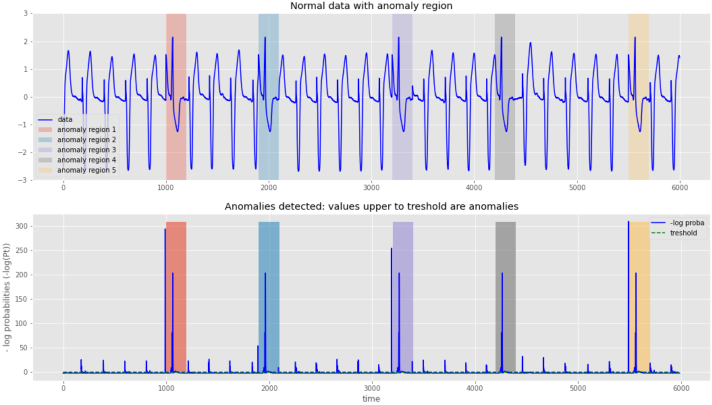

Figure: ECG dataset with many anomalous intervals

Results : ECG / Respiration datasets

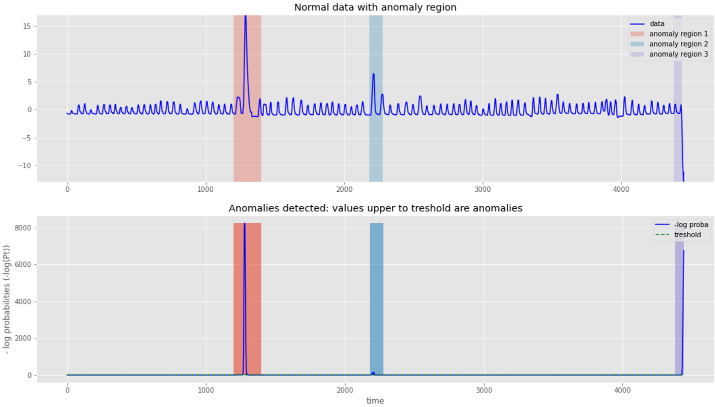

Figure: Repiration dataset

Results : recap

Dataset	R2 score	MSE	Recall	Precision	$F_{0.1}$ score
ECG (one anomaly block)	0.99	0.003	0.03	0.34	0.32
ECG (Many anomaly blocks)	0.99	0.045	0.08	0.55	0.53
Respiration	0.98	0.008	0.02	1.0	0.68

Table: Metrics for performance evaluation

Results : Synthetic data

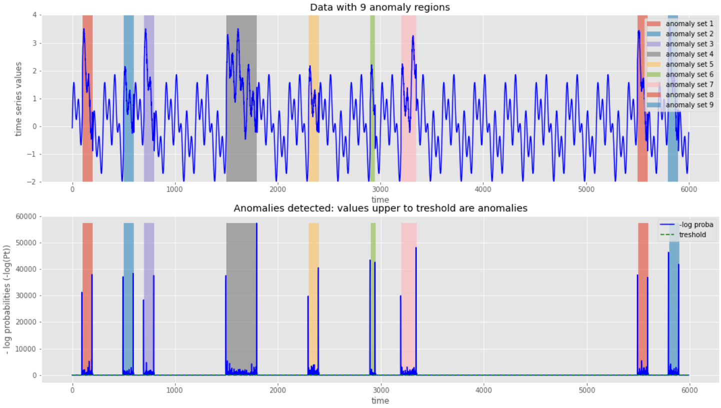

Figure: data without any kind of noise

Results : synthetic data, Noise effect

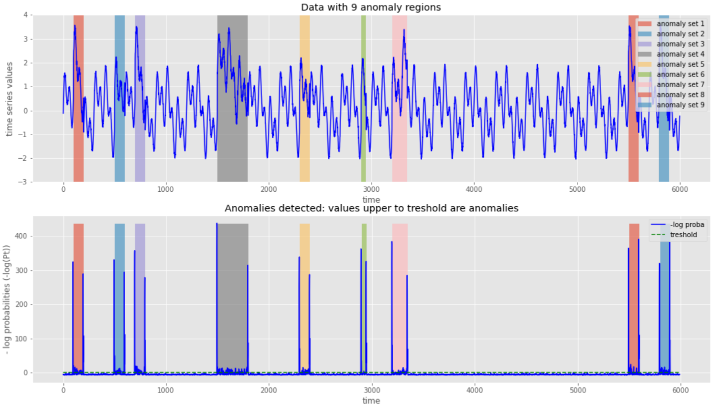

Figure: data with small noise, $\sigma = 0.05$

Results : synthetic data, Noise effect

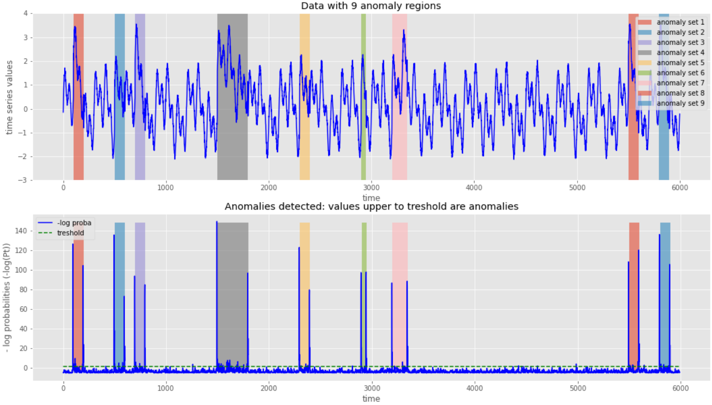

Figure: data with small noise, $\sigma = 0.1$

Results : synthetic data, Noise effect

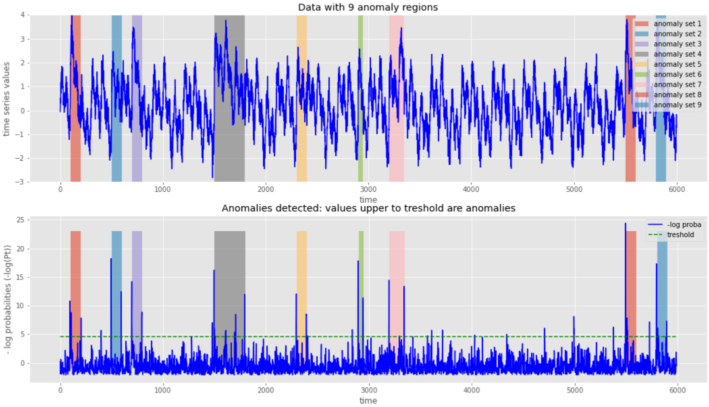

Figure: data with small noise, $\sigma = 0.4$

Results : synthetic data, Noise effect

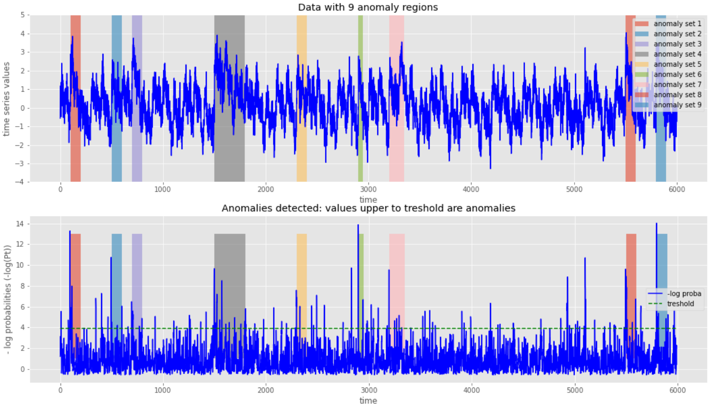

Figure: data with small noise, $\sigma = 0.8$

Results : recap

Noise level	R2 score	MSE	Recall	Precision	$F_{0.1}$ score
$\sigma = 0.0$ (no noise)	0.997	0.002	0.19	1.0	0.95
$\sigma = 0.05$	0.98	0.006	0.08	0.99	0.88
$\sigma = 0.2$	0.93	0.058	0.01	0.96	0.57
$\sigma = 0.4$	0.83	0.15	0.01	0.74	0.51
$\sigma = 0.8$	0.58	0.38	0.03	0.59	0.51

Table: Metrics for performance evaluation

Conclusion

Recap:

1. Advantages
   - Model based : LSTMs are extremely good models for modeling and understanding time series features.
   - Rich method : LSTMs can be trained on h steps in the past to predict l steps in the future. $\Longrightarrow$ Model complex time series, multi-variate time series ...
   - The more we have anomalies in the data, better the method gets !
2. Drawbacks
   - Supervised regression based method : Model should be trained again if we change time series.
   - No Statistical guarantees !
   - Sensitivity to noise !

The End

References

Pankaj Malhotra, Lovekesh Vig, Gautam Shroff, Puneet Agarwal
Long Short Term Memory Networks for Anomaly Detection in Time Series
https://www.researchgate.net/publication/304782562_Long_Short_Term_Memory_Networks_for_Anomaly_Detection_in_Time_Series
2015