NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

School of Computer Science and Engineering
College of Engineering

Deep Learning based Anomaly Detection in Multivariate Time-Series Data

1. Background

Anomaly detection, also known as outlier detection, is the identification of rare observations that do not conform to the expected pattern of a given group. Anomaly detection is an essential task in multiple domains, such as: intrusion detection, fraud detection, monitored engine signal anomaly detection, etc. Common attributes among these domains are that the data are multivariate time-series data.

2. Challenges

Multiple challenges exist in the research:

- Constantly evolving boundary between normal and anomalous behaviour
- Anomaly diversities and lack of labelled data
- Data monitored for anomaly detection are usually heterogeneous, noisy and high-dimensional
- Results generated from the models require interpretability

3. Research Objective

Our research focuses on the anomaly detection in multivariate time-series data with the use of unsupervised deep learning based artificial neural networks, in designing efficient anomaly detection algorithms.

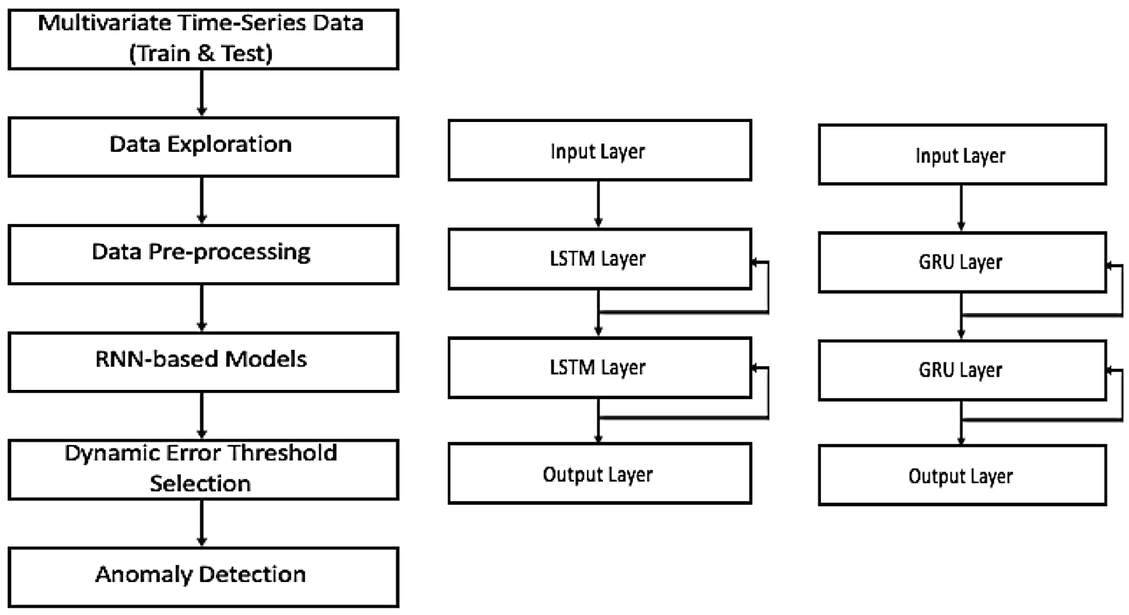
4. Datasets

The datasets used in this project are Soil Moisture Active Passive Satellite (SMAP) and Mars Science Laboratory rover (MSL). Both SMAP and MSL are public datasets from NASA.

Dataset name	Number of entities	Number of dimensions	Training set size	Testing set size	Anomaly ratio (%)
SMAP	55	25	135183	427617	13.13
MSL	27	55	58317	73729	10.72

5. Methodology

We propose an anomaly detection framework as shown in (a). RNN-based models include stacked LSTM model (b1), stacked GRU model (b2). The prediction errors are calculated: $e^{(t)} = |y^{(t)} - \hat{y}^{(t)}|$. Errors are smoothed through the formula in (c1). Dynamic error threshold ε is determined by (c2). The anomaly score is calculated in (c3) to assign to each anomalous sequence indicating the severity of the anomaly. Anomaly pruning is experimented to reduce false positives. As shown in (c4), if $d^{(i)} < 13\%$, the anomaly sequence will be reclassified as normal.



(a) (b1) (b2)

$$EMA_t = \frac{x_t + (1-\alpha)x_{t-1} + (1-\alpha)^2 x_{t-2} + \ldots + (1-\alpha)^t x_0}{1 + (1-\alpha) + (1-\alpha)^2 + \ldots + (1-\alpha)^t}$$

(c1)

$$\epsilon = argmax(\epsilon) = \frac{\Delta\mu(\mathbf{e_s})/\mu(\mathbf{e_s})) + (\Delta\sigma(\mathbf{e_s})/\sigma(\mathbf{e_s})}{|\mathbf{e}_a| + |\mathbf{E}_{seq}|^2}$$

where:

$$\Delta\mu(\mathbf{e_s}) = \mu(\mathbf{e_s}) - \mu(\{e_s \in \mathbf{e_s} | e_s < \epsilon\})$$
$$\Delta\sigma(\mathbf{e_s}) = \sigma(\mathbf{e_s}) - \sigma(\{e_s \in \mathbf{e_s} | e_s < \epsilon\})$$
$$\mathbf{e}_a = \{e_s \in \mathbf{e_s} | e_s > \epsilon\}$$
$$\mathbf{E}_{seq} = \text{continuous sequences of } e_a \in \mathbf{e}_a$$

(c2)

$$s^{(i)} = \frac{max(\mathbf{e}_{seq}^{(i)}) - argmax(\epsilon)}{\mu(\mathbf{e_s}) + \sigma(\mathbf{e_s})}$$

(c3)

$$d^{(i)} = (e_{max}^{(i-1)} - e_{max}^{(i)})/e_{max}^{(i-1)}$$

(c4)

6. Result

The result is shown in table (d) below in terms of precision, recall and $F_{0.5}$ score. Stacked LSTM models demonstrate better overall performance than stacked GRU models. Error pruning effectively minimize false positives at the cost of false negatives.

Methods	SMAP			MSL		
	Precision	Recall	$F_{0.5}$ score	Precision	Recall	$F_{0.5}$ score
Stacked LSTM w/ Pruning (p=0.13)	**0.855**	0.855	**0.855**	**0.962**	**0.694**	**0.893**
Stacked LSTM w/out Pruning (p=0)	0.485	**0.928**	0.536	0.862	**0.694**	0.822
Stacked GRU w/ Pruning (p=0.13)	0.831	0.855	0.836	0.870	0.556	0.781
Stacked GRU w/out Pruning (p=0)	0.430	0.879	0.479	0.852	0.639	0.799

Student: Zeng Jinpo

Supervisor: Assoc Prof A S Madhukumar