Anomaly Detection from Sensor Data

Mohammad Poul Doust, Andrei Mardale, Laetitia Couge, Bognan Etienne Ekpo

Msc Machine Learning and Data Mining (MLDM)
Jean Monnet University

January 13th, 2020

Introduction

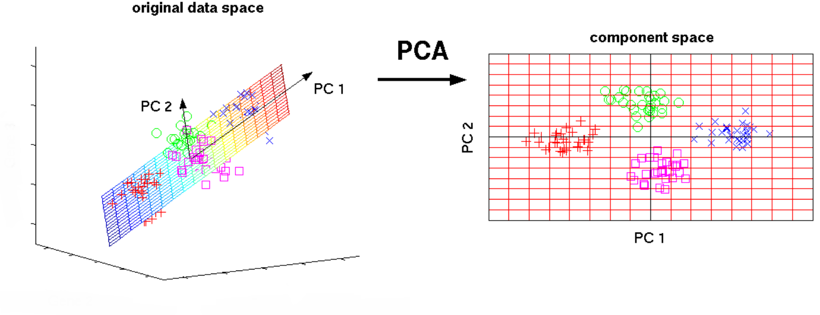
Problem Context: Similar to previous pitch:

- Training: 1677 sequences, All normal !
- Validation: 600 sequences
- Test: ~ 2000 sequences
- Each sequence ~ 60000 points

Goal: Use this data to learn the true representation of normal samples and use it to detect Abnormalities.

Challenges

- Large feature vectors ~ 60 k
- No Labels in training
- Learn the true representation of Normal
- Anomalies in the final test data different from validation.
- When an anomaly is detected on one flight, all sequences of this flight are labelled as abnormal
- ,,,

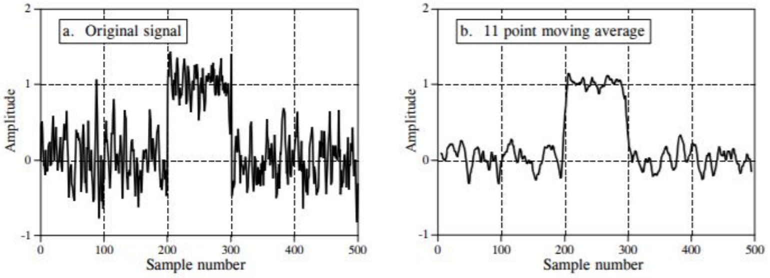


Effectively reduces the dimension. BUT, Loses temporal information and it is Linear

Solutions: Large Dimensionality

- **Moving Average**: Sliding window that consider the average of each window as a representative.

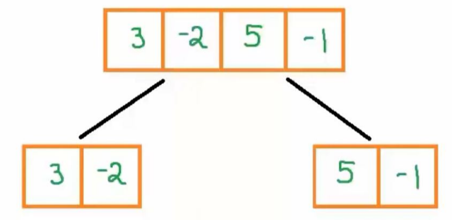
- Window of size 1000 (one second) $\Rightarrow$ the new signal of size 60.



Effectively reduces the dimension. BUT, Choosing window size is tricky and average suffers from compensation

Solutions: Large Dimensionality

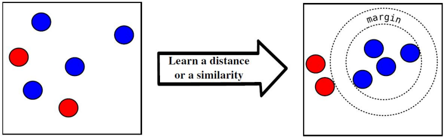
- **Creating new dataset**: Splitting each sequence into K new sub sequences and treat each one as a new dataset instance
- ⇒ Increase dataset size, Decrease feature vectors.
- ⇒ Less prone to overfitting.
- K = 12 ⇒ each sample will give 12 new instances, each of size 5000



Effectively reduces the dimension and give more samples.
BUT, needs to tune K and extra processing when classifying

Solutions: Large Dimensionality, Learn Representation

- **Metric Learning**: Learn a metric that assigns small (resp. large) distance to pairs of examples that are semantically similar (resp. dissimilar)[1].
- Lower Dimension
- Better Representation
- Needs some Label !
- Linear, but could be Kernelized !



Results

Rank	Team	University	F1-Score	Precision	Recall	Method
2	The-h-star	Universite Jean Monnet Saint Etienne	0.99	0.98	1.0	Metric Learning
40	The-h-star	Universite Jean Monnet Saint Etienne	0.93	1.0	0.88	Autoencoders + OCSVM + Splitting
80	The-h-star	Universite Jean Monnet Saint Etienne	0.89	1.0	0.80	PCA + OCSVM

Table 1: Ranking at Nov. 29, 2019

Conclusion

- Anomaly detection could be very tricky
- Dimensionality Reduction highly impact model performance and accuracy
- Metric Learning is effective tool to learn better data representation, but needs labels
- Autoencoders also effective to learn better representation (No need for labels)