NPL
National Physical Laboratory

Machine Learning for Cavitation Detection

David Sinden

Medical Ultrasound | Chemical, Medical and Environment Sciences | National Physical Laboratory
david.sinden@npl.co.uk

New Technologies for Clinical and Preclinical Research | 5 December 2018

Contents

NPL National Physical Laboratory

Quote 1

“Artifical Intelligence is one of the most important things humanity is working on. It is more profound than, I dunno, electricity or fire”

Google CEO Sundar Pichai

Why machine learning?

NPL
National Physical Laboratory

The hype is because the recent availability of

- large, validated, labelled datasets;
- improved hardware;
- clever parameter constraints;
- advances in optimization algorithms;
- enabled by more open sharing of stable, reliable code

have contributed to the rapid successes of machine learning

But what is machine learning?

NPL National Physical Laboratory

Machine learning is different from classical statistical methods in that, in general

"statistics draws population inferences from a sample, and machine learning finds generalizable predictive patterns."[1]

Its utility is from an ability to make reliable predictions beyond the scope of a data set[2].

[1] Bzdok et al. *Statistics versus machine learning*, Nature Meth., (2018) 15, pp. 233-234

[2] Pathak et al. *Using machine learning to replicate chaotic attractors and calculate Lyapunov exponents from data*, Chaos (2017) 27(12), pp. 121102.

Is it relevent and useful?

NPL National Physical Laboratory

The portability, cost, rapid acquisition of data and ubiquity of diagnostic ultrasound mean that it is at the forefront of image classification and analysis.

Prediction aims at forecasting unobserved outcomes, but without necessarily knowing how a system fully works. It looks at correlation, not causality.

Indeed, one of the advantages of machine learning is that it can be effective when the data is acquired without a carefully controlled experimental design, or perhaps more appropriately, when there is variablity between experimental setups (or detection methods), or in the presence of complicated nonlinear interactions.

Applications to cavitation

NPL National Physical Laboratory

Two types of task:

- catagorization

or

- quantification

Two types of data:

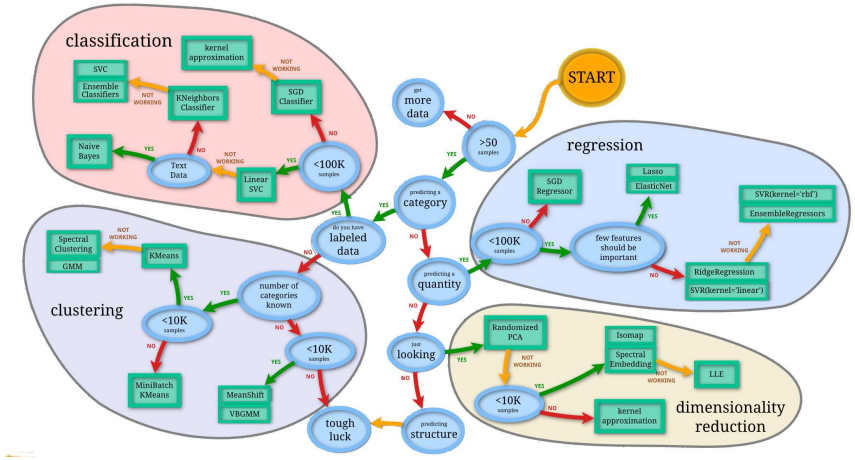
- labelled, i.e. supervised learning

or

- unlabelled, i.e. unsupervised learning

Guide to algorithm selection

National Physical Laboratory



Nomenclature

The data is divided into training and test datasets: the algorithm learns from the training dataset and then is assessed against the test dataset.

A useful rule-of-thumb is that there should be at least five times as many examples in a training data set as there are features.

The data sets are simply matrices comprised of values from the features. The features are in-part derived from the experimental setup as well as the data acquired.

Automatic feature extraction

NPL
National Physical Laboratory

Preliminary results to automatically ascertain features from both spectral images and time-series data did not yield clear definitions.

A possible reason is that the “images” are very diverse

The approach often imposes constraints on image type and format — such as number of pixels, which restricts representation

Features

NPL National Physical Laboratory

Feature selection and extraction is the most important part of any end-to-end machine learning approach.

A large number of features provide a more robust algorithm

One interpretation is defining the features of a data set amounts to codifying intuition.

This requires domain-specific expertise, so may not be a task for a data scientist . . .

Features

NPL National Physical Laboratory

Features should be

- indepedent
- informative

and

- discriminating.

What are the features of a cavitation data set?

Features: measured data

NPL National Physical Laboratory

Firstly, post-processed data

- Integrated broadband[3]
- Subharmonics[4]
- Width of half-harmonic[5]
- Kurtosis
- Sum-of-harmonics[6]
- Crest[7]
- Wavelet-based methods[7]
- ...

[3] W-S. Chen et al *Ultrasound Med. Biol.* 2003 29, pp. 725—737
[4] Song et al, *J. Acoust. Soc. Am.*, 2017, 141(3), pp. EL216–EL221
[5] N. Segebarth et al. *Chem. Phy. Chem.* 2001, 2(8), pp. 536–538
[6] E. Lyka et al. *J. Acoust. Soc. Am.*, 2016, 140(1), pp. 741–754
[7] S. R. Haqshenas & N. Saffari, *J. Phys. Conf. Ser.*, 2015, 581, pp. 012004

Features: experimental setup

NPL National Physical Laboratory

Secondly, measurement set-up

- Frequency,
- Transducer geometry,
- Exposure duration,
- Medium,
- Sensor type,
- Filtering,
- …

Features: labelled data

National Physical Laboratory

Thirdly, for supervised learning, a label is placed within the feature set

For labelled data a challenge is ascertaining whether cavitation has occured without the using the features — needs gold standard data sets, in which cavitation activity is known without using the features, such as optical, thermal methods . . .

Feature space

NPL National Physical Laboratory

The upper dimension of the feature space is $\sim$25.

Most data sets provides over a 100 waveforms

Dimension reduction techniques, such as generalised principal-component analysis can be applied, which will create a smaller number of ranked, independent quantities, based on a linear combination of features.

What can be done?

NPL National Physical Laboratory

At this stage, not attempting to predict any metric of cavitation dose given a set of parameters, but classifying a simple (binary) choice as to whether cavitation has occurred or not.

If the data is labelled this is straight-forward. It is simple within the same framework to classify whether inertial, non-inertial or non-cavitation (trinary) has occured.

Why hasn't this been done before?

NPL National Physical Laboratory

There are two articles of interest[8,9]:

Lee et al. used a technique to predict degree of cell membrane disruption at 24 kHz: defines features as: broadband, half harmonic, subharmonics as well as exposure time, peak positive pressure, pulse length and duty cycle.

Gregg et al. looked at fatigue and failure in turbine blades, based on: RMS, crest, peak value and kurtosis from four differing types of sensors and a number of locations.

[8] Prediction of Ultrasound-Mediated Disrution of Cell Membranes Using Machine Learning Techniques and Statistical Analysis of Acoustic Spectra, *IEEE Trans Biomed. Eng.* 51(1) 2004, pp. 82-89

[9] Feature Selection for Monitoring Erosive Cavitation on a Hydroturbine, *Int J. Prognostics & Health Man.* 2017 3

Algorithm selection

NPL National Physical Laboratory

The natural choise was to use a support vector machine.

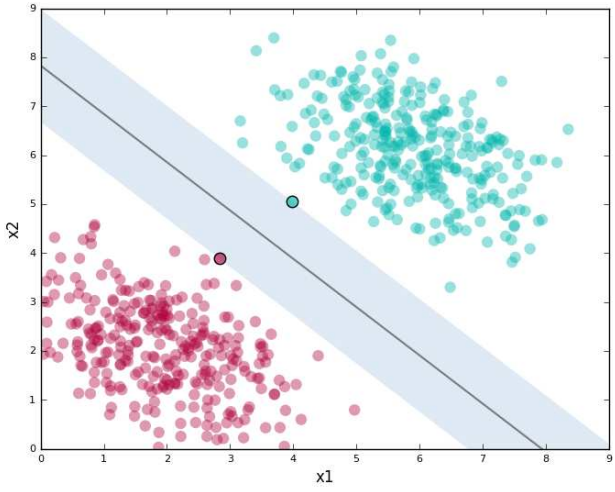
This algorithm draws a surface which separates, by maximizing the distance from the values nearest the surface for each class

- The algorithm assumes data can be mislabelled
- It will also permit exposures to be classed as non-cavitating when it may be labelled cavitating
- Furthermore, weights can be placed on data whoses integrity is doubted
- Outliners can be identified and discarded within the minimization process
- A linear or nonlinear boundary can be defined

A random forest approach would also be possible.

Support Vector Machines

NPL National Physical Laboratory



Data sets

Data sets from:

- Leeds
- ICR
- Oxford
- NPL

were collected. Data was over a range of frequencies, powers, exposures durations, etc.

Preprocessing

NPL National Physical Laboratory

For each data set a `python` module was created which would extract the features and create the feature matrices.

A small piece of code then applies the machine learning algorithm.

An issue was that different data sets would be emphasize different features: for example to (i) maximize signal-to-noise in integrated broadband data and (ii) reduce risk of saturation in digitisation, a high-pass filter may have been applied. This would suppress half-harmonic data.

Another factor was that the NPL data was only in the frequency domain.

Classification

National Physical Laboratory

Within data sets, over a large number of tests, ~ 50

Data set	Likelihood of Success	False Positives	False Negatives
1	85%	11%	4%
2	84%	11%	5%
3	89%	9%	2%
4	90%	8%	2%

with 10:1 training to test datasets using `svm.svc` algorithm with a linear kernel from the `scikit-learn` package

Classification

NPL National Physical Laboratory

A universal algorithm for long duration exposures

Data set	Likelihood of Success	False Positives	False Negatives
—	60%	19%	21%

Quote 2

NPL
National Physical Laboratory

“Most of machine learning is the software equivalent of banging on the side of the TV set until it works, so don’t be discouraged if you have trouble seeing an underlying theory behind all your tweaking”

Peter Warden, data scientist and member of Google’s TensorFlow team

Recipe

NPL National Physical Laboratory

- Is the aim to predict a quantity or catagorize events? Choose appropriate algorithm.
  - Which quantities will be predicted?
  - If categorization how is data labelled? Are there gold standards?
- Acquire data with features in mind. Acquire sufficient data in relation to features, algorithm and computational capabilities
- Standardise and normalize data
- Divide data between training and test data sets
- Train algorithm
- Evaluate accuracy

Conclusions 1

NPL National Physical Laboratory

Criteria for reporting how data is acquired is necessary for intercomparisons and generalisations of the algorithms to multiple laboratories or exposure conditions.

Any experimental setup seeks to gain the best desired information, which may mean that some features are not measured, reducing the dimension of the feature space. A problem is that if data is dominated by thresholds of broadband noise, then algorithms essentially learns how to do signal processing

Intercomparison of data is challenging, but possible. In-house classification is demonstrated.

Reliably using labelled data is challenging, as the labels may be derived from metrics used in feature space.

Future work

NPL National Physical Laboratory

- Extend code to better handle sparse and incomplete data-sets, i.e. when some features are not measured.
- Explore simulation parameter space to extend code to better handle a single data set from mutliple setups
- Extend code to agnostically process data, regardless of material and exposure conditions

Finally 1

National Physical Laboratory

As may have been mentioned -

Senior Research Physicist in HIFU NHS AfC: Band 8a

Main area: research and development
Grade: NHS AfC: Band 8a
Contract: Fixed term 2 years
Hours: 37.5 hours per week
Job ref: 196-BRC1230

Employer: Guy's and St Thomas' NHS Foundation Trust
Employer type: NHS
Site: St Thomas
Town: London/Teddington
Salary: £49,077 - £56,632 p.a. inclusive of HCA
Closing: 04/01/2019 23:59
Interview date: 15/01/2019

Finally 2

NPL National Physical Laboratory

AMUM 2019 — Advancing Measurements in Ultrasound in Medicine
will be held at NPL in 14-17 May 2019.
More details will be announced shortly.

Thank you for your attention

NPL National Physical Laboratory

Thanks for your time, and thanks to colleagues at ICR, Oxford and Leeds for data, as well as colleagues at NPL, and ThUNDDAR and the NMS for funding

Any questions?

david.sinden@npl.co.uk

@david_sinden

NPL
National Physical Laboratory

Department for
Business, Energy
& Industrial Strategy

FUNDED BY BEIS

The National Physical Laboratory is operated by NPL Management Ltd, a wholly-owned company of the Department for Business, Energy and Industrial Strategy (BEIS).