Machine Learning in support of expert quality control

Community

Quality Control

SG-IQuOD Meeting 2018
Oostende, Belgium

Guilherme Castelão
castelao@ucsd.edu
Inst. Dev. Group - Scripps
Spurious data are inevitable; thus, quality control is a requirement.

Modern times demand efficient QC response:

- **Low latency** - Realtime data
- **Large volumes** of data - Historical review and reproducibility

An **Ideal QC** would be as **fast** as an automatic QC and **flexible** with **top quality** as a manual QC
Which ML technique to use?

- Unsupervised learning: Better representation and simplification of the problem
- Requirements for supervised learning:
Machine Learning classification

NN
SVM
Clusters

IQuOD 2015
Machine Learning classification

M.L.

NN

SVM

Clusters

…
Machine Learning classification

Spurious data can be anything; hence, they're not representative.

M.L.

A good classifier doesn't mean a good predictor for new data.

Real samples must obey physical laws.

NN SVM Clustering ...

A good classifier doesn't mean a good predictor for new data.

Spurious data can be anything; hence, they're not representative.
Which ML technique to use?

- Unsupervised learning: Better representation and simplification of the problem
- Requirements for supervised learning:
  - Calibrate with a representative dataset
  - QC is an unbalanced problem (luckily $<< 1\%$)
- Anomaly Detection
Quality Control Criteria

- Feasible limits
- Climatology test (z-score threshold)
- Independent multiple feature climatology (Gronell 2008)
Quality Control Criteria

❖ Feasible limits
❖ Climatology test (z-score threshold)
❖ Independent multiple feature climatology (Gronell 2008)

Questions:
❖ Could we obtain some information from the other tests?
❖ Is a Gaussian distribution a good assumption?
More degrees of freedom means more flexibility on decision making

Castelão 2015, 2016
Anomaly Detection

- Criterion based on valid data, so intrinsically:
  - No problem with unprecedented spurious measurement
  - No problem with unbalancing of bad samples
- Recognize what is anomalous
  - Spurious measurements
  - Real rare events
Expert QC user interface

https://expertqc.castelao.net
Community QC

- Users
- Personal calibration
- CoTeDe release
- Community calibration
- Automatic QC
- Flags dataset

https://expertqc.castelao.net
Products

- Close the cycle by returning the calibrated CoTeDe to the community
- Provide open access to the flagging dataset, so it can be used to calibrate other techniques
Challenges

- Time for development. Maintenance is minimized by automation in the cloud
- Scientific development is a small fraction of the requirements to keep such system running
- Computationally expensive
Future work

❖ Recruit, engage, and learn from QC experts (in progress);
❖ Employ advanced/alternative techniques for specific problems:
  ❖ Wire break/Hit bottom for XBT (Rebecca Cowley- CSIRO);
  ❖ Pattern identification (Ruth Musgrave - WHOI)
❖ Anonymize human flags and make them open access;
❖ Propose a procedure for (cross-)validation of experts (trainees);
❖ Training/educational: Teach new quality controllers to identify patterns;
CoTe De l’eau

- Open Source and Free Software
- Customizable with preset (GTSPP, EuroGOOS, …)
- Run in parallel for efficiency
- Easy to extend available tests
- Easy to integrate (AutoQC uses CoTeDe)
- But it is not perfect. Please, let me know your complains so I can improve it
- http://cotede.castelao.net