Data mining experience at HIPA

Jochem Snuverink, Andreas Adelmann
Contents

• Data mining
• HIPA
• Data mining at HIPA
• Summary
Per capita consumption of mozzarella cheese correlates with Civil engineering doctorates awarded
• Most common types of data mining\(^1\)
  - **Anomaly detection** (outlier/change/deviation detection)
  - **Dependencies** (finding relationships between variables)
  - **Clustering**
  - **Classification**
  - **Regression** (finding a model function)
  - **Summarisation** (compactifying dataset, visualisation)

\(^1\) "From Data Mining to Knowledge Discovery in Databases" – Fayyad, Piatetsky-Shapiro
I used to think correlation implied causation.

Then I took a statistics class. Now I don't.

Sounds like the class helped.

Well, maybe.
Six phases (according to CRISP-DM)

1. Problem understanding
2. Data understanding
3. Data preparation
4. Modeling ← here Machine Learning could be used
5. Evaluation
6. Deployment
Problem understanding: HIPA

Ring Cyclotron 590 MeV
2.2mA / 1.3MW
diameter: 15m

meson production targets

proton therapie center
[250MeV sc. cyclotron]

SINQ
spallation source

dimensions:
120 x 220m²
ML ideas

• Interlocks prediction and prevention (see talk Andreas)
  - Benefit: Increased uptime, reduced damage and activation
  - Why ML: Many different interlocks, large amount of archived data for training
  - Which type: Supervised learning with time-step memory (RNN)

• Minimise beam losses
  - Benefit: Less activation, possibly reach higher current
  - Why ML: Many tuning knobs in complicated machine
  - Which type: Reinforcement learning (Policy Gradients)

• Target spot size optimisation
  - Benefit: Improved machine protection and neutron yield
  - Why ML: Direct measurement (optical monitor) relatively slow. Predict and interpret the optical image based on machine diagnostics
  - Which type: Supervised learning with image recognition / prediction(?)

• Stable isotope production
  - Benefit: Improved isotope production and less operator supervision
  - Why ML: Product quality only afterwards. No good predictive model
  - Which type: Supervised learning with some sort of data reduction
Data understanding

- About 80 GB data / year
  - 30+ years of data
- 3 machine databases with different set of variables:
  - Short term (<10 days)
  - Medium term (<10 months)
  - Long term
- Variables are logged differently
  - monitored versus different fixed rates
- Interlock database
- Isotope production quality (excel sheet)
• Select list of variables: magnet currents, diagnostics (BPMs, loss monitors), etc.
  – problem dependent
  – for interlock study only ring cyclotron and high energy line
• Get data from archive with command line tool:
  - Variables missing ₋ -> adjust list, about 90 left
  - Simple usage puts a new line for every parameter change
    – Unmanageable for more than a few variables
    – Fixed data rate of 10 Hz (in interpolated values)
  - Crashes when trying to limit significant values
  - 120MB / day (binary format), 1e7 samples
  - Missing values -> remove whole sample
• Combine interlock data
  - add interlocks to closest timestamp (are we sure clocks synchronised?)
Data selection

• Clean variables
  - variables not changed significantly (from physics point of view)
  - remove variables that have correlation higher than 0.95 with another
    – will probably not add much to a model
• Clean samples
  - Remove bad machine conditions
    – machine off or ramping: current < 1300 uA
Visualisation – scatter plot matrix

Probably not so useful
Visualisation – box and whisker plot

Median

25-75%

IQR

CR3V:IST:2 Cavity voltage

Outliers

1.5*IQR

Median

Whiskers and Outliers

Cavity voltage

867.8
867.6
867.4
867.2
867.0
866.8
866.6
866.4
866.2

Cavity voltage
Visualisation – correlation matrix

- Extraction and RF in ring cyclotron
- Trim coils in ring cyclotron
- Phase and loss monitors in ring cyclotron
- Loss monitors in targets beamline
- Interlocks
Modeling – binary classifier
see talk Andreas

density plot
some discrimination power
but many false positives
• Network weights

• As expected the variables with significant weight correspond to those with high correlation.
• Simple methodology for data mining
• Personal experience on HIPA data shown
  - Data preparation step most tricky
    - Discuss with controls group how this can be improved
  - Data normalisation needed for ML
• Some simple visualisation plots that can guide for large amounts of data
• Simple regression model
  - Reduce false positive rate
  - Add predictive power (RNN)