REMOTE DIAGNOSTICS, MACHINE LEARNING AND DATA COLLECTION FOR AUTOMATED HIGH THROUGHPUT MASS SPECTROMETRY

Emmy M Hoyes[1], Thomas C. Smallwood[1], Nicola Johnston[1], Richard C Chapman[1], Allen Caswell[2]

[1]Waters Corporation, Wilmslow, UK, [2]Waters Corporation, Milford, USA

Waters

THE SCIENCE OF WHAT'S POSSIBLE.®

INTRODUCTION

Machine learning is the collective name for techniques that automatically trains models to recognize patterns in the data associated with different states or outcomes. It is very powerful as it can deal with a large number of variables but typically requires many datasets to be able to train a reliable model. Speech recognition and spam filters are a couple of typical applications using machine learning. By applying it to our mass spectrometers we hope to develop a system that can tell us when and what maintenance and repair is required. This strategy allows for remote detection of issues and means that we know what needs to be done before an engineer arrives on site which minimizes instrument downtime.

In this poster two different studies are described showing the potential of and challenges associated with applying machine learning to mass spectrometry instrumentation. The first proof of concept study utilised the built in health monitoring system of the QDa to see if the correct state of the instrument could be detected using the available settings and readbacks from the instrument. The second study involves production engineers labelling data from quadrupole time-of-flight mass spectrometers (XEVO G2-XS) as they were making sure the instruments pass specification.

QDA PROOF OF CONCEPT PROJECT

The built in health monitoring system of the QDa mass spectrometer was modified to allow for rapid and automatic prototyping for the machine learning project as described in Figure 1. The data included all readbacks and settings available from the instrument together with a label describing what that had triggered the event. The study included 8 different instrument either present in-house or at the BioHub at Alderley Park, Macclesfield. Gathering data for the QDa project was relatively easy as the labelling of the data was done automatically by the health monitoring system and it included various instrument states rather than just instrument failure which provided data set at a higher frequency.

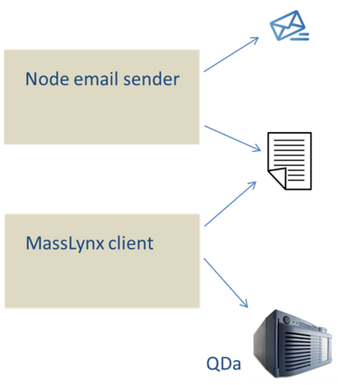

Figure 1: MassLynx was modified to collect and send to file a full set of readings and settings when a diagnostic event was triggered. A simple node.js script monitored the file and forwarded a copy of it via email to a dedicated email address.

PROCESSING

A python application was developed to retrieve the datafiles from the emails and parse them into a csv file suitable for machine learning. A total of 814 datasets were included in the final analysis of which each contained the values from 178 readbacks or settings (down from 719 by removing any constant variables and highly correlated variables) describing 32 different health states. Machine learning typically requires a large number of replicates for each state to be accurately classified. With the limited replicates for certain states we included only states with 10 replicates or more, which is a low number but still gave very good results. The data set was split 70:30 into a training set and a testing set to allow for checking the accuracy of the machine learning model.

MATLAB and the python scikit-learn library was tested for their machine learning capability and both showed similar results. Several algorithms were investigated such as Naïve-Bayes, k-nearest neighbors algorithm (kNN) and random forest to name a few. The data was modified according to the algorithm used which meant that sometimes only continuous variables (such as voltages) or categorical variables (such as interlock info) or a combination of both were used to train and test the model.

RESULTS

Figure 2 shows the results summarized as a confusion matrix from the random forest algorithm which was the best performing algorithm. The purple diagonal line highlights the high proportion of successes where the actual class and the predicted class coincide. It demonstrates that certain states are easier to accurately detect than others but that overall the method is pretty good at accurately classifying the different states.

Just 15 states covered 94% of the events received from QDa instruments. On average the algorithm correctly classified the state of the instrument 87% of the time if the state of the instrument was in any of these 15 states. The remaining 6 % of the data was spread over 17 different relatively rare states which meant that no individual state had enough data points associated with them to be able to train the model to accurately classify hem. With more data points this would be possible. It is very useful to know which states might be easily confused and the frequency of misclassifications. It is also quite possible that two or more statuses are true at any one time. This is tricky as the machine learning will only ever return one answer and only when that issue or state is no longer true might a different one resurface.

The random forest algorithm was the best performing algorithm possibly due to it performing very well on data that consists of a mixture of continuous and categorical data.

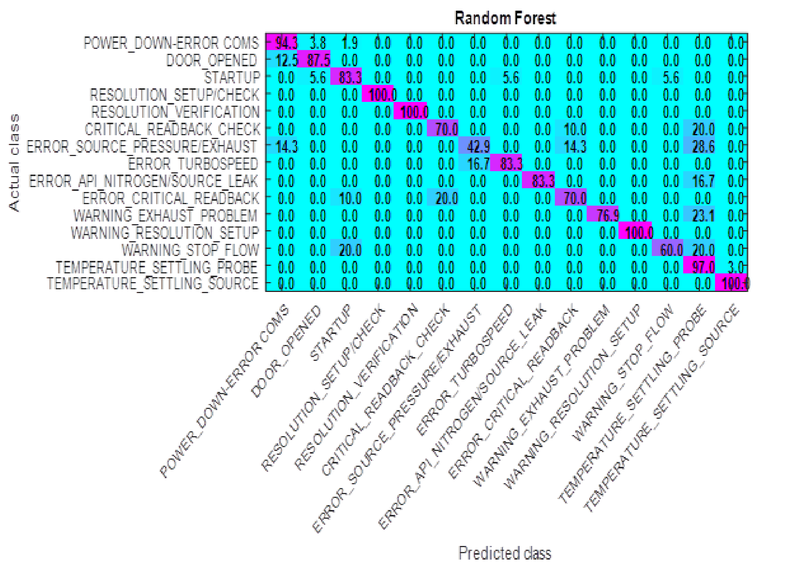

Figure 2: Confusion matrix showing the ability of the random forest algorithm to predict the state of the instrument. The diagonal line shows the percentage of accuracy for each state.

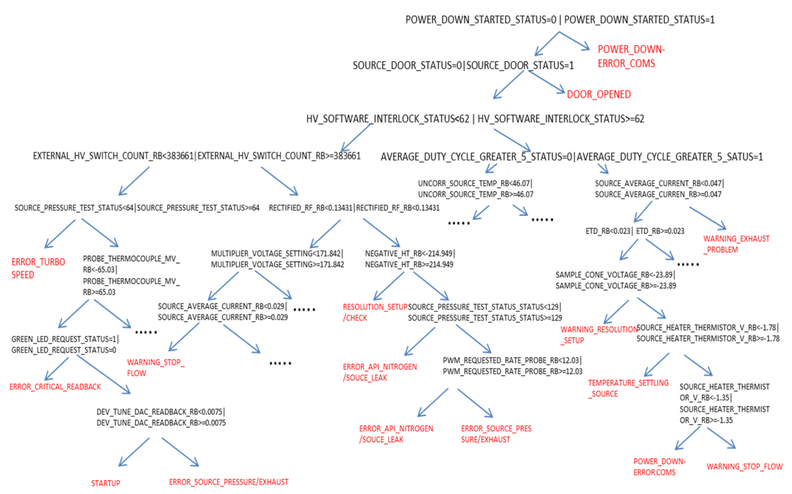

Figure 3: Example of a single decision tree showing the logic applied by the algorithm. The random forest consists of an ensemble of several hundred decision trees chosen based on a random selection of the variables in the data which increases the accuracy compared to just a single tree.

CHALLENGES OF DATA COLLECTION

Machine learning improves with the quantity and the quality of the available data. Some sectors will have accumulated such data over the years such as financial trading patterns and medical research data. The maintenance and repair of scientific instrumentation does not generally have records associated with it which include enough information for machine learning to be applied retrospectively. Thus, collection of this data will need to be done and for it to succeed it needs to fit in with the workflow of existing procedures and be consistent.

The next step was to implement the machine learning in a more realistic setting, with the relatively high throughput Xevo G2-XS instruments going through in house release testing chosen as the target.

XEVO G2-XS PROJECT

A software application was developed and deployed to all instruments undergoing final tests after being built to allow for a streamlined and user-friendly collection of all settings and readbacks by the engineers when a fault was detected on the instrument. As well as instrument properties additional information such as serial number, symptom, sample type and a typical spectrum was also recorded. After the fault had been fixed a second set of readbacks and settings were recorded together with information on how the instrument had been fixed including part numbers. The data was then automatically copied to a server to be available for processing. After the engineer has collected a set of readings and provided a symptom, the application presents a webpage with fixes recently associated with the reported symptom and where enough data has been provided the machine learning model can highlight the most likely fix based on the data submitted.

The data collected from the Xevo G2XS has very high accuracy as it is collected in a very reproducible way using a script and labeled by trained engineers. Our main issue was the high variability in reported symptoms, with each symptom receiving only a small number of reports, thereby not generating a high enough number of data points for fast implementation of machine learning.

INCREASING THE NUMBER OF DATASETS

As more data leads to better results we have tried to maximize the data collection at each failure event by adding in extra collection cycles whereby the script first gathers the usual readbacks and settings before varying a commonly used settings (e.g. capillary voltage) within a defined range, and recollecting the same data using the new setting values. This is repeated 20 times using a randomly selected values each time. By doing this it is possible to create more datasets where each dataset will have some variation similar to what would be seen between instruments. This is extremely difficult to do well as the data produced must be identical to what would have been produced in the field, otherwise the machine learning model will be worthless.

This approach will certainly not negate the need for many occurrences of each fault but will make it possible to more quickly identify predictors that are suitable for including in the model and for preliminary reports to be generated.

The result from a preliminary model is shown in Figure 4 where 15 selected parameters have been included in the model to predict 5 different conditions.

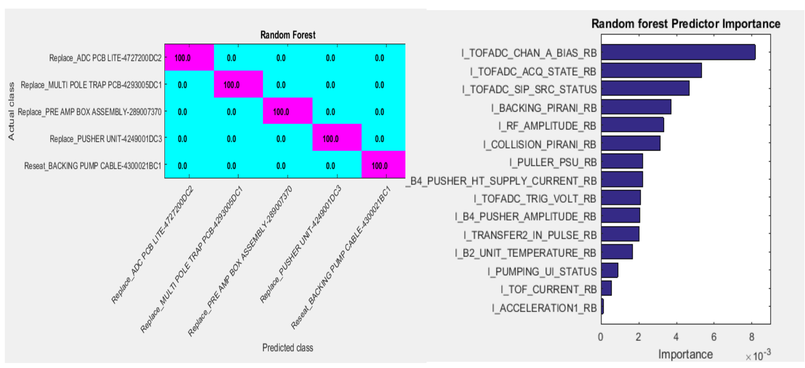

Figure 4: Preliminary results of 5 selected XEVO G2-XS repairs showing that despite only 25 occurrences in total (20 replicates added of each occurrence to provide more data) it is possible to use machine learning to classify the repair required in this particular case based on 15 parameters collected from the instrument for which the importance is shown above.

CONCLUSIONS

- Machine learning shows promise as a way of automatically detecting the state of MS instrumentation.
- Work is underway to use machine learning to provide a tool for trouble shooting in a production environment with further scope to extend to instruments in the field.
- To be able to generate a reliable model you need many datasets for each condition which is difficult and time consuming to collect.
- Attempts to speed up the data collection process are fraught with biases as by nature it will not be truly representative of the real thing but can aid in predictor selection.
- One advantage is that the model learns to detect the most common conditions first, while it can continue to learn as more data is accumulated on more rare conditions.

TO DOWNLOAD A COPY OF THIS POSTER , VISIT WWW.WATERS.COM/POSTERS