Machine Learning Engineer Nanodegree

Capstone Proposal

Brian Mello
June 2019

Proposal

Domain Background

In the field of instrumental analytical chemistry, there are many techniques for identifying unknown substances. This field is primarily researched for the detection of explosives, drugs and chemical weapons, however, there are many other applications for it. In the healthcare field, researchers are using it to detect cancer and diseases.

Chromatography is a method by which a substance is separated into different components and split across time. For example, if you sample a mixture that is made up of multiple substances you could separate those substances and detect them individually with chromatography. One method of chromatography is using what is called a column. This device is what will separate out the components of a mixture, however, it is extremely slow. It takes minutes to work.

Mass spectrometry is the technique of identifying the chemical structure of a substance by separating ions into differing mass and charge. The three essential components of a mass spectrometer are the ion source, the mass analyzer, and the detector. The ionizer ionizes the sampled substance. The analyzer sorts and separates the ions by mass and charge. The detector measures or counts the separated ions. Ions can be displayed as a histogram of mass-to-charge ratio(m/z) or can be displayed across time to see line curves where the peaks would be where the max quantity of ions were detected. There are many different techniques for each component in performing the identification of substances.

The most popular analytical technique today is ion-mobility spectrometry (IMS). This technique separates and identifies ionized molecules based on their mobility within a carrier gas. It is extremely fast and takes milliseconds to achieve a result, but it is less sensitive than other techniques. This technique is very popular because you can make an IMS device for relative low cost compared to other techniques and the device can be small enough to be hand-held.

The final technique that we will discuss is triple quadrupole mass spectrometry (TQMS). This is a tandem mass spectrometry technique, which just means it has two analyzers. The components in order are the ion source, a quadrupole mass analyzer, a quadrupole that acts as a collision cell to fragment the molecules entering it, a second analyzer that analyzes the resulting fragments, and finally

the detector. The quadrupole works by creating an electro-magnetic field that separates the ions and makes them follow a trajectory based on their mass-to-charge ratio (m/z). This technique in theory is the most sensitive and will achieve results in seconds. These devices tend to be very expensive and large. This is the device a team and I are working on.

The above techniques discussed can all be combined to solve problems depending on the application. There are trade-offs that must be made for each technique. Cost and weight are always a major factor. In some cases, the science is not well understood.

Problem Statement

My team and I are currently working on a triple quad mass spectrometer that is cheaper and smaller so that we can address new markets and applications that TQMS was unable to address previously. Our current instrument displays mass-to-charge ratios over time. We have in the past used peak thresholds to determine what is a detected substance. This technique only gives us an accuracy of 40% with a very high rate of false positives. We are trying to achieve an accuracy of 90% with no more than a 2% false positive rate. We have talked about adding some filtering techniques, but there will be a tradeoff in time and cost. We need to complete our analysis in under 10 seconds. Ideally, we can solve our problem with purely algorithms.

Datasets and Inputs

The datasets that will be used in this project were generated from collected samples from our instrument. The instrument was sent out for testing and 12 different substances were tested. The data files we got back are the results from that testing. The datasets are generated from these data files. The datasets have been modified to abstract out any sensitive details such as the substance name and mass pairs. Most importantly the intensities are all generated to mimic the shape of the collected data. The data has also been filtered to remove any malformed data because of a hardware or any other error. There is no proprietary data associated with this project. The model that will be built will need to be re-trained on the actual proprietary dataset to have it work with our instrument. The generated data should be more than adequate in evaluating a model. In most cases, I have between 50 and 80 samples for each substance I am testing for. I realize that this may not be enough, but I am also trying to gauge how many samples will be needed if extending out the substance library. If a compound is performing poorly from lack of samples I will remove it from the test.

Data file columns and example:

mass_pair_id	sample_id	comment	substrate	detection	association	peak_height
0	21321	Blank	None	None	[1,2]	1000.0

...

peak_width	peak_area	peak_position	timestep1	...	timestep_n
5.2	3000.2	5.1	0.0	...	100.5

Each data file consists of multiple components. First, there will be a mass pair transition id. A mass pair transition consists of an ion charge (+ or -), parent mass, a daughter mass, and collision energy i.e. +123->456(78). The ion charge is from the ion source, the parent mass is from the first quad, the daughter mass is from the third quad and the collision energy is applied at the second quad. Instead of seeing this transition you will see a number 0 to n-1 where n is the total quantity of specified transitions (i.e. n=52). After the id there will be a sample id associated to the dataset, a comment field which will specify if another substance was combined the tested for substance, and what substrate it was sampled on. Substrate could have the value direct, or a harvest code like Perf5. Direct means the substance was inserted into the instrument through a syringe. Direct should be the most stable result. If the substance was harvested off material Perf5 that means the substance was applied to the material and then swiped off with one of our swabs. Theoretically when a substance is harvested it should measure lower ions than direct because you may not of collected the entire amount of the substance. After the substrate field, there will be detection field and an association field. The detection field will have an array of numbers specifying what compounds are detected within that dataset. The association field will specify what mass pairs are associated to which compounds according to our chemist, for example mass pair 1 is associated to compounds [1,3]. After the associated field there will be a height, width, area, and position of the mass pair peak. These values are acquired by applying a smoothing filter to reduce the noise of the signal. Finally, there will be a time series of intensities over a specified number of time steps or scans (i.e. 23). For now, the scan count is fixed to 23, however, in the future it would be better to have scan count be variable in case we want to stop early. So, for each data file you would have that structure per row times the amount of transitions for example 52x33.

One potential detail we are not addressing is potentially using a different filter for our mass pair signals. We are using a gaussian filter with a value of sigma equal to 1. This allows us to smooth the data. If we raised sigma I'm not sure if we could gain more consistent results. There are varying opinions on this topic and for simplicity we will not address it for now.

Solution Statement

In the past, using thresholds was not enough information to correctly classify substances. One problem was it removed a lot of important details of the signals that were acquired. If mass pair 1 is in compound 4 and 5 when compound 4 is tested mass pair 1's shape is going to look a lot different than when compound 5 is tested. I plan to use a recurrent or convolutional network to give me a probability of what compounds exist in a series of mass pair signals. I should receive good results because the scientists and I can tell when we have a detection based on the signal shape. I do not think we can rely entirely on shape. Peak properties are important as well. I might need to create an algorithm that my CNN/RNN can feed its results into a parametric or non-parametric algorithm that incorporates the peak properties.

Benchmark Model

The model can only be benchmarked against the previous solution which yields a total accuracy of 40%. Minimum peak height was the only threshold where it was manually set based on internal lab testing. for example, compound 1 could have a minimum height threshold of 1600 ion counts. If the peak signal was below this amount it was deemed noise and ignored. If it was above it would be part of some additional logic that required all associated mass pairs to be above their limits to raise a substance detection alert. Below is most of a confusion matrix, but TNR is missing.

Compound ID	TPR	FPR	FNR
Compound 3	5.00%	4.55%	95.00%
Compound 4	30.34%	52.81%	32.58%
Compound 7	0.00%	87.50%	12.50%
Compound 8	13.33%	20.00%	66.67%
Compound 10	12.31%	7.69%	86.15%
Compound 13	0.00%	21.21%	84.85%
Compound 14	5.00%	0.00%	95.00%
Compound 15	20.37%	12.96%	68.52%
Compound 18	59.65%	28.95%	14.91%
Compound 19	20.00%	70.00%	50.00%
Compound 21	72.90%	26.17%	0.93%
Compound 22	70.27%	17.12%	15.32%

Accuracy: ~40%

Evaluation Metrics

To evaluate our models, we should be using a weighted accuracy metric such as fbeta_score and a confusion matrix to see our false positive rate. According to our requirements we could miss a detection 10% of the time if we have a false positive rate below 2%. Since it is more important to be precise than have high recall we should have our beta be a value of 0.5.

Project Design

The very first thing we should do is some exploratory data analysis. I first need to assess all the peak properties. Ideally, I would like to use a parametric solution with the peak properties as input because it would require less sampling results to determine what parameters make up a positive detection. I do believe that there may be some strong correlation between some of the peak properties. These features would have to be removed from our model to not negatively influence the training of the model. Based on the results of a scatter matrix, I would also need to assess if the usable parameters are gaussian or not. If they are not gaussian, I will need to apply a natural log or box-cox transformation to them to try and make their distribution gaussian. If they are still not gaussian, I will have to use a neural network to try and find a solution.

Based on how poorly our benchmark model performs, I am assuming a more complex solution will be needed than just using the peak properties. I would like to incorporate the shape of our mass pair signal to determine if that is something that can be used to narrow down. As previously stated, I would like to use a convolutional or recurrent network to pull features out of the time series that can be assessed. My plan would be to use a max scaler on the data, so all values are between 0 and 1. This way the height and width do not matter, just the shape. I might need to add some dropout to the network, so I do not overfit on the data, but based on the variability of the signal I feel like I will not need to. I will use 60% of the data for training, 20% for validation, and 20% for testing. I will need to plot the test results so that I can determine the best threshold value. As much as I would like to have one CNN/RNN for determining my multi-label problem, I think I will have to split it into one network per associated mass pairs. For example, the compound 0 CNN/RNN would take as input the scans of only its associated mass pairs [1,2,3]. This ensures that if there is any consistent chemical noise it will not be incorporated into the model. The only way we know if a mass pair is associated is if the mass pair is validated by a chemist who has run an analysis on it.

Depending on how good, the shape data works out I might need to incorporate some of the peak properties into the model. One issue I must address is the

scaling of the peak properties. Let's look at height for an example. There is no theoretical max for what the peak height could be. I could use the max of an individual scan series like what I am doing with the shape analysis, but it could create a problem if there are large peaks of alternative mass pairs that dwarf the relative height of my peak. This could cause my model to miss a positive detection. I think the only thing I could do is to apply a standard scaler to a mass pair's individual properties so that other mass pair peak properties do not negatively influence others.