University of Zurich[UZH]

Zurich Institute of Forensic Medicine

Deep Learning – Creation of an Artificial Neural Network for the Prediction of Data Generated by High Resolution Mass Spectrometry with Data Independent Acquisition

Elmiger Marco[1], Dobay Akos[2], Ebert Lars[3], Kraemer Thomas[1]

[1]Department of Forensic Pharmacology and Toxicology, Zurich Institute of Forensic Medicine, University of Zurich, Zurich, Switzerland

[2]Department of Forensic Genetics, Zurich Institute of Forensic Medicine, University of Zurich, Zurich, Switzerland

[3]Department of Forensic Imaging/Virtopsy, Zurich Institute of Forensic Medicine, University of Zurich, Zurich, Switzerland

1. Introduction

General unknown screening (GUS) in biological matrices becomes more and more crucial in forensic toxicology with an ever growing number of NPS entering the drug market. LC coupled to high resolution quadrupole time-of-flight mass spectrometry (LC-HR-QToF) provides a suitable analytical platform to meet this challenge. While data-dependent acquisition (DDA) approaches are still widely used, the greater possibilities provided by data independent acquisition (DIA) approaches are more promising for GUS. The disadvantage of DIA approaches is the huge amount of data produced which has to be dealt with. The promising field of deep learning offers new possibilities where neural networks can be trained to classify big amounts of data. The aim was to exemplify such an approach for HR-MS DIA files by comparing different deep learning approaches (KNIME, Keras and TensorFlow).

How does the model learn?

The main objective in a deep learning model is to reduce the loss function's value. The loss is calculated on a training set and its interpretation is how well the model is doing. It is a summation of the errors (usually residual sum of squares) made for each example in a training set. Ideally, one would expect the reduction of loss after each, or several, epoch(s). An epoch is a single pass through the entire training set (Fig. 1).

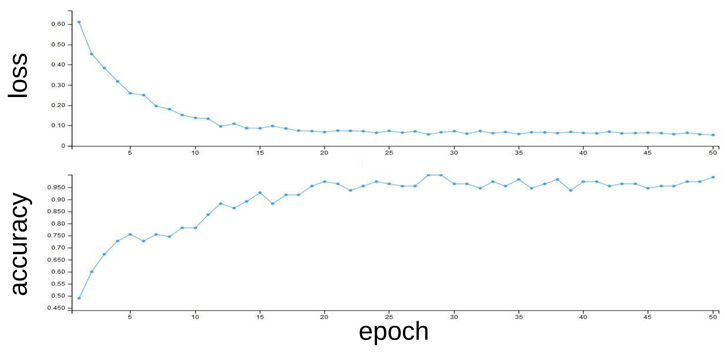

Fig. 1: Example of the progress of loss and accuracy over several epochs in a deep learning approach

Feeding the neural network

To train the neural network, a train set was composed including 50 blank, 50 cocaine and 50 zolpidem blood samples. With this train set, the network should learn and find similarities and differences between the groups.

The test set includes several blood samples (blank, cocaine, zolpidem and combinations of the latter two with other substances) which the neural network did not see before.

Sample number	Blank	Cocaine	Zolpidem
1	0.87	0.11	0.02
2	0.14	0.83	0.04
3	0.32	0.57	0.11
4	0.84	0.14	0.02
5	0.89	0.08	0.03
6	0.07	0.88	0.05
7	0.89	0.09	0.03
8	0.43	0.53	0.04
9	0.03	0.02	0.94
10	0.69	0.07	0.24
11	0.08	0.85	0.07

Sample number	Blank	Cocaine	Zolpidem
1	1	0	0
2	0	1	0
3	0	1	0
4	1	0	0
5	1	0	0
6	0	1	0
7	1	0	0
8	1	0	0
9	0	0	1
10	1	0	0
11	0	1	0

Fig. 2: Example of a prediction table generated by the deep learning approach (left) compared to the true values (right). Blue boxes show final category decision (left) and true category (right)

Conclusion

Using HR-MS DIA generates a huge amount of data which has to be dealt with. Deep learning by neural network approaches offers new possibilities to handle big amounts of data and classify them. In this project it was possible to prepare HR-MS DIA data files via an R script to make them available for deep learning approaches. Furthermore, it shows the possibility of sample type prediction by a deep learning approach after learning only by a training set. The accuracy and precision for the prediction via different deep learning approaches of both the test and validation set were in a good range. Differences between the different deep learning approaches were very little and a decision between them would be more of a handling preference type. This project shows that deep learning promises a huge potential for the handling of big data generated by DIA with high resolution mass spectrometry

3. Results

- Different neural networks were built in KNIME and Keras by creating a workflow or writing an R script, respectively. After the neural networks learned on the train set similarities and differences between the groups, they were able to do a prediction of the test set including samples that the neural network did not see before (Fig. 2)
- With the confusion matrix it was possible to calculate the accuracy and precision of the validation and test sets via the values predicted by the different neural network approaches (Fig. 3).
- Both deep learning approaches reached ideal and very similar values for accuracy and precision with the test set.

Fig. 3: Comparison of accuracy and precision values for the test set generated by different deep learning approaches using several HR-MS DIA data. Accuracy and precision was calculated due to the confusion matrix and the equations (TP + TN)/(TP + FP + FN + TN)x100% and TP/(TP + FP)x100%, respectively

2. Methods

- **Data collection** via HR-MS combined with DIA (TOFMS/ SWATH) on a Sciex TripleTOF 6600 with a 5 minute method using a RP-C18 column in a Dionex UltiMate 3000
- **Samples:** Blood samples from authentic cases (blank, cocaine, zolpidem, combinations of the latter two)

- **Data preparation** was done by first converting the data via Proteowizard from the Sciex .wiff to the open data format .mzML
- Data cleaning using an R script, where the 20 most intense peaks per cycle time (described by RT, m/z and intensity) were chosen
- Structure was generated via the same R script by transposing the cleaned data to a table. Each row describes one data file by many rows.

- **Data analysis:** Comparing Deep Learning models on different platforms (KMIME, Keras and TensorFlow). The structure of the models had three dense layers between the input and output layers with output units of 1000, 100, 10 for the dense layers and three for the output layer.

Contact

Marco Elmiger, marco.elmiger@irm.uzh.ch, www.irm.uzh.ch