Bachelor Thesis Presentation

Stefan Moser, 20.1.2020, Start 10:50, End 11:25

Protocol by Günter Klambauer.

- What is the reason for your choice of the classifier? Why do you chose xgboost and not a neural network?
  - *Xgboost is a simple gradient boosting technique that is known to work well for tabular data as given here. The advantage over neural network is that they are easier to train and less hyperparameter search has to be done.*

- What do you do with compounds unknown toxicity?
  - *As I performed single-task training, these values could be discarded without problems.*

- What types of data sets did you look at?
  - *I worked with the ChEMBL database, a data set of several millions of small molecules. I worked with the Tox21 dataset, too, which contains outcomes of toxicity tests.*

- What is the meaning of “screens” or “screen” in you presentation?
  - *Screens are ways to scan a database, like virtual screening. Screens can discard or select molecules. In machine learning this means generating a bit vector. And provide an algorithm for labeling.*

- Do I understand correctly that you first generate molecules with an LSTM and then check by the xgboost models?
  - *Yes, this is correct. I use an LSTM to generate a lot of molecules. I then filter this molecules with the models. The models from xgboost that predict toxic or not toxic.*

- What is a “valid” molecule?
  - *In this work I consider it valid when the SMILES is correct. SMILES is the line-syntax for small molecules. There are brackets and ring-openings and closings. These have to be correct. If not closed, this would not be a valid SMILES.*

- Are there other ways to generate molecules with a computational technique?
  - *Any type of RNN could work. However, regular RNNs would have a disadvantage over LSTMs. RNNs can be used.*

- Are there other ways to check for reasonable or good small molecules?
  - *It would be good to see when they are represented in the data set. It is good to just provide a generator that produces molecules. We can use a base model for transfer learning.*

- What was the overall goal of your study?
  - *The goal was to provide a data set of drug-like molecules. These molecules should not be toxic. We filtered out toxic molecules. This data set is a good starting point for drug discovery projects. This can save time.*

- Is there a way to adjust the “false negative rate”?

    - *I adjusted for the ratio of toxic and non-toxic molecules. I balanced the data sets. Perhaps by hyperparameter search or weighting the classes.*

- Why did you suggest to use all potential SMILES characters even though not present in the data set?
    - *This would have led to a more diverse data set. All characters are defined in the SMILES syntax, so there is no need to extract them from the training data set.*