Enhancing Toxicological Testing through Machine Learning

M. Leone, G. Macchi, M. Vicentini, M. Baity-Jesi, K. Schirmer

Abstract

A short documentation to get started with the project.

1 Introduction

We want to be able to predict the effect of untested chemicals on species that were already tested on (*across chemical exploration*), and to predict the effect of known chemicals on untested species (*across species exploration*).

The standard procedure for across-chemical exploration is based upon exploring the similarity in composition of different mixtures.

2 Databases

- Chemical structure: Chemspider
  `https://www.chemspider.com`
- Chemical structure: Dragon
  `https://chm.kode-solutions.net/products_dragon.php`
- Chemical testing on organisms: Ecotox.
  `https://cfpub.epa.gov/ecotox/search.cfm`
- *In vitro* chemical testing: Toxcast
  `https://www.epa.gov/chemical-research/exploring-toxcast-data-downloadable-data`
- Combination of several:
  EnviroTox
  `https://envirotoxdatabase.org`
  ChlamyChem
  `http://chlamychem.utoronto.ca/ChlamyChem/download.php`

2.1 EcoTox database

The database is pretty large: 11'756 Chemicals; 12'906 Species; 49'153 References; 952'634 Results. However, we won't use all the features, and we don't know whether they are informative enough.

Some of the features:

- CAS number identifies the chemical. In a second moment we can use it to add features from each chemical from the Chemspider database.

- Organism (to be grouped e.g. for different species of water fleas or on a higher aggregation level, such as crustacae).
- Exposure length (duration): in most experiments it's 4 days. There is no direct way to infer the relationship between duration and effects, and it does imply a significant change.
- Effect type – called endpoint (there are many in the database: acute tox, reproduction, growth)
- Effect concentration. This is usually the target. If the endpoint is LC50 (lethal concentration 50%), this is the required concentration of toxine to kill half of the population. The column name is `CONC. TYPE: CONC. (STD)`, and the values require some cleaning.
- Organism life stage. You want to discard embryos.
- Exposure type: this is how the concentration of chemical is kept constant. Can be interesting at a later stage. Static: they fill the tank at the beginning and assume nothing changes; flow through: the concentration is kept constant through a continuous flow in the tank; renewal: some shady procedure which isn't trusted.

How to filter the data We will start studying mortality.
Effect: Mortality.
Endpoint: LC50.
Species group: Fish.

The target is the concentration required to kill half of the population, `CONC. TYPE: CONC. (STD)`. This regression problem can become a classification problem if we divide the targets in mortal/non-mortal by setting a concentration threshold $\mathcal{T}$. We can start with $\mathcal{T} = 1 \mathrm{mg}/\ell$.

Legend of the columns: `https://cfpub.epa.gov/ecotox/help.cfm?sub=wi-definitions`.

2.2 Chemspider database

The Chemspider database will be used to get the features of the chemicals.

We might want to use the KOW from this database, which is an indicator of the hydrophobicity of the compound. Hydrophobic means that it gets attached easily to fat, and gets absorbed better.

3 Concrete objectives

A non-exhaustive list of possible questions we might want to answer:

- Can we identify the effect of tested chemicals on new species?
- Can we identify species sensitivity to untested chemicals?
- Predict the exposure concentration at which a chemical causes effects (endpoints) such as mortality, growth, reproductive failure, etc...

- Can we identify families of chemicals/species with a major
- Can we use mammalian data (which is very abundant) to make predictions on aquatic organisms?
- Can we use *in vitro* data to make predictions on mammalians?

4 First Steps

1. Exploration of the dataset. Start with EcoTox: Basic statistical analysis of the dataset. Produce the typical fluff that gives us an idea of what the databases look like (number of data, incomplete/redundant columns, outliers, ...).
2. Identify a test harness (an initial simple database on which we can work on in a relevant way), a performance measure (e.g. loss), and train-validation-test sets within the harness (with cross validation) for one of the following tasks:
   - Train models that identify the impact of old chemicals on new species
   - Train models that identify the impact of new chemicals on old species
   - Apply clustering methods to identify structure in the data
3. Clean the data (categorical encodings, feature generation, feature selection)
4. Setup pipeline: train a dummy model. Also, be aware of what the results would be with the worst possible model (i.e. random predictions).
5. spot-check possible models. Choose $\sim 5 - 10$ candidates and run their vanilla version through the test harness.
6. Check predictions. Which features are the most important? How does this correspond with the field experts expectations? Is there any test that we can perform to make sure that our results are sound?

Once these steps are done, we are in good shape.