Tesi di Laurea – BOZZA

Simone D'Ambrosi

October 21, 2020

1 Data Structure

The ECOTOXicology knowledgebase (ECOTOX) is a source for locating single chemical toxicity data for aquatic life, terrestrial plants and wildlife. I focus my attention on the aquatic kingdom to predict the effect of chemicals on fishes that were already tested on.

From ECOTOX Database I extract three different sources regarding informations about fishes, chemicals and the conditions of laboratory experiments.

Most of experiments consist on a tank where fishes live and freely swim, then a chemical is introduced. The main differences among the experiments regard the total chemical concentration, how many times it is injected, how long fishes manage to live in these conditions, which fishes species and its type are tested on. A good example in order to better understand how these experiments work is to consider a tank filled with clean water where some fishes are present. Then, the clean water is replaced by a contaminated one in order to annotate how long fishes can survive. In addition, even test's control may change depending on its type.

Subsequently, the obtained results are collected and stored in the ECOTOX database with accurate information about which effect is intended.

Another fundamental fact is the endpoint of the experiment which involves the different quantification of an observed effect obtained through statistics or other means of calculation for the express purpose of comparing equivalent effects. (source ECOTOX website)

Moreover, experiments are conducted many times since the results of a single try may not be typical for the population. Hence, repeated experiments require particular attention in the manner in which they are merged together. Often, the same experiments could lead to different results making it less reliable than others. This effect have some benefits as well as drawbacks. If on the one hand we can merge same experiments and fill the gaps due to not available or not reported data, on the other hand they can rely on far different results making the joining ambigous. Therefore, the mean of results is not a good choice since it significantly suffers from strange values or outliers. A more appropriate statistic for this kind of problem may be the median, but an excellent idea is to consider a confidence measure for repeated experiments. It would have to work in this way: suppose we have 2 different experiments, A and B, both repeated 5 times and results may be either H or K. If experiments A obtains 3 times as result H and 2 times K, and experiments B obtains 5 times result H and 0 times K, then experiments B must be more trustworthy than experiments A. In some sense, each repeated experiments have to consider as well the "static" aggregation of the results as the "dynamic" aggregation in order to have a better trust in the data.

For the sake of simplicity, I suppose to use the median, hence considering all the experiments worthy of trust.

ECOTOX database is strongly influenced by the authors of the experiments themselves that cause a significant lack of information due to not reported field of the database. Therefore, a screening step is required since most of the fields turn out to be not available or, as previously stated, not reported depending on the author's choice.

The results of this first phase point out specific characteristics about the name of the fish species, the environment in which they live (salt or fresh water), what chemical is tested and some relevant information about the duration and the type of the experiment.

1.1 Species data

Tested organisms may be studied by a genetic point of view to understand how fishes, and in general animals, are related. Every organism is classified by a taxonomic name which refers unambiguously to it. ECOTOX database referred to tested fishes not only through a key value but also with the taxonomic name concerning the kingdom (Aquatic, Terrestrial, Plants, Wildlife), the phylum division and so on. Figure 1 describes the Animal Taxonomic Classification and its structure.

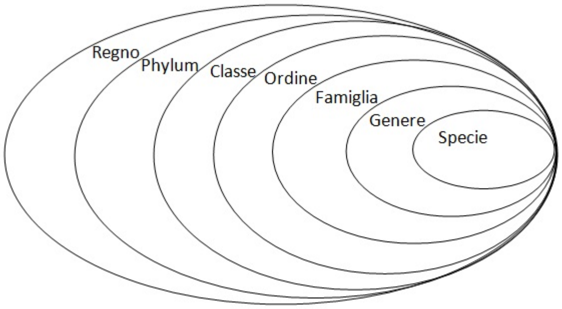

Figure 1: Taxonomic Classification, da mettere in inglese

In order to study the structure of evolutionary relationships among biological entities from a mathematical point of view, often species or genes, it's possible to make a phylogenetic analysis of the data. Phylogenetics is more than the simple study of species, indeed it examines the history of organism in its evolution through time.

The first step of the analysis starts with handling missing data. Authors can have forgotten or partially reported the name of the tested species. In any case, due to the impossibility of any imputation I drop out of the dataset all species that have one or more not available/reported field. The same procedure is been applied to the environment where a specific species lives, but, in addition, to make the dataset more consistent I decide to take off all fishes that are not neither freshwater nor saltwater.

Once ECOTOX database have been cleaned up so that the remaining tested animals are exclusively Fishes, it turns out the presence of 580 fishes species, 416 freshwater and 164 saltwater. By a taxonomic point of view those fishes are divided into 3 classes, 29 taxonomic orders, 106 families, 322 genus and 514 species. The configuration of the so-called phylogenetic tree may be built in Python using ETE Toolkit, a Python framework for the analysis and visualization of trees.

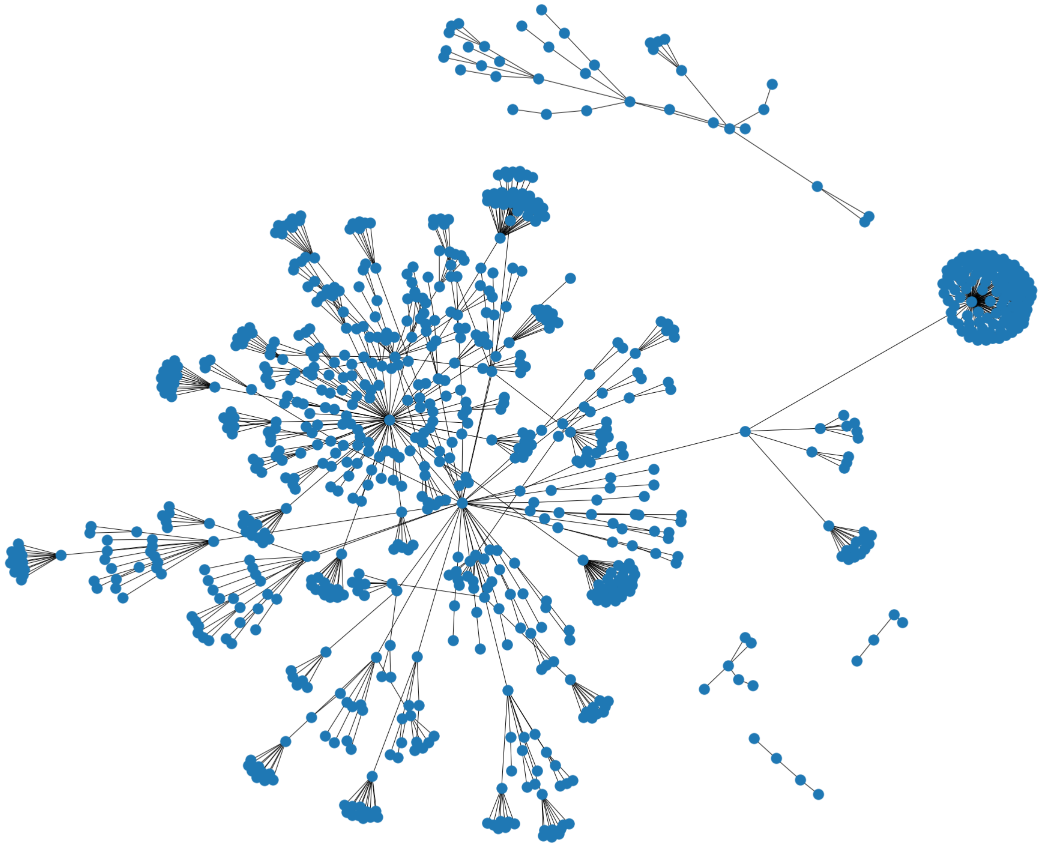

Figure 2: Phyogenetic tree, ne ho uno di gran lunga migliore ma è in pdf

1.2 Tests data

One of the goal of this work is to predict the lethal concentration of chemicals able to kill half of the tested population using Machine Learning algorithms. With the aim of reaching this purpose, the screening phase of ECOTOX data have to take into account which Endpoint and Effect are required.

For Endpoint, ECOTOX refers to the objective of experiments such as the lethal concentration able to kill 1%, 10%, 20%, 50% and so on of the tested population (LC1/LC10/LC20/LC50...) or the concentration able to inhibite organisms (IC).

For Effect, instead, ECOTOX refers to some change in attitude, reproducibility, growth or death.

However, an example can help to approach to this terminology. As previously said, one of the goal is to predict the lethal concentration of a specific chemicals able to kill half of the tested fishes. Clearly, I want to predict both the toxicity of a chemical and the concentration to assume that chemical to be toxic. In this case, I focus my attention on the endpoint LC50 and EC50 that stand, respectively, for Lethal Concentration and Effective Concentration, and on the effect MOR that stands for Mortality.

Later, I'll study also other effects in order to apply more sophisticated tools as Data Fusion RASAR.

Back to how experiments are made, most of trials are developed in laboratory and a minimum part in the real wildlife. In laboratory, the condition are more controlled and the impact of chemicals is under control. It may vary for different conditions such as how and how many times the chemical is injected in the tank.

Authors of testing experiments refers as *exposure type* the manner in which the chemical is fed, as *application frequency unit* the application rate of the compound, for example, *una tantum* is designed as *X*, X times for 20 minutes as*X for 20 MI* and so on.

As regarding the *exposure type*, to make it more practical, an example is needed. Fishes are in the tank, swimming in clear water. Thus, chemical have to be given them, but how can author do that? Actually, in several different ways. ECOTOX exposure type codes provide plenty informations about it. It's possible to spray chemical onto the skin of the fish or via intradermal injection, and still, through a diet or spread in the environment, granular or food, and again, by changing the water with polluted water or fishes can be pull out of the water and the chemical can be injected by air. The way chemical is given to the fishes have a huge impact on the results of the experiments. Missing data are removed since imputing means affect substantially the experiments.

Once the dose has been given and the period of time of the trail is over, author must check the desired effect. Also this information can vary. It can be either satisfactory or unsatisfactory result, it can resolve to be insufficient to gain the intended effect, the author can choose to check the effect on only one fish or make multiple types of control or eventually, if the author did not report the control type, I assume it to be Unknown, to avoid to introduce biases due to imputation.

1.3 Chemical data

About the procedure of the experiments, the only element left to make it a real part of ECOTOX database is concerning with the administered substances.

Every chemical substance described in the open scientific literature, including organic and inorganic compounds, minerals, isotopes, alloys and non-structurable materials, is named by a unique numerical identifier assigned by the Chemical Abstracts Service (CAS). CAS Numbers are generally a serial numbers, so they do not contain any information about the structures themselves. It has no inherent meaning but is assigned in sequential, increasing order when the substance is identified by CAS scientists for inclusion in the CAS Registry database.

In other words, scientists refers in a fancy and usable fashion to chemical compounds simply using CAS Registry Numbers.

Of course, the purpose of this work is to determine the effective toxicity, and so to predict the effect, of untested chemical on fishes that were already tested on (across chemical exploration). In most of the real cases, it's not possible to know which chemical has been diluted in the water, but is possible to extract some relevant information about the structure of that chemical. Thus, the CAS Number is no more a such relevant detail.

For this reason, I use the package CirPy to "translate" CAS Number to SMILES. The simplified molecular-input line-entry system (SMILES) is an ASCII string describing, as far as possible, the structure of a chemical. From SMILES is possible to extract several extremely useful features such as the atom number and the molecular weight, the number of (single, double and triple) bonds, the lipophilicity or rather a parameter linking solubility, membrane permeability, and hence drug absorption, and many others.

One of the most important descriptors of a chemical compound is the Pubchem2d Fingerprints. A fingerprint is an ordered list of binary (1/0) bits. Each bit represents a Boolean determination of, or test for, the presence of, for example, an element count, a type of ring system, atom pairing, atom environment (nearest neighbors), etc., in a chemical structure.

This list of binary bits is composed by 881 position, hence, determining the presence or the absence of 881 different characteristic.

Finally, a list of all features is provided:

- About fish species:
  - Class
  - Taxonomic Order
  - Family
  - Genus
  - Species
  - Media Type: freshwater or saltwater
- About chemicals:
  - CAS Number
  - SMILES:
    - Atom number
    - Alone atom number
    - Bonds number
    - Double bonds number
    - Triple bonds number
    - Ring number
    - Molecular weight
    - Morgan Density
    - Lipophylicity (aka LogP)

    - Number of ”OH” group
  - Pubchem2d Fingerprints

- About tests:

  - Period of observation: it consists on how many time fishes are subjected to the treatment. It can vary from hours to months
  - Time unit of the period of observation: all the tests are turned into hours
  - Concentration of injected chemical: each author can choose chemical's concentration to be injected and report this quantity in the unit measure he/she prefers.
  - Concentration unit measures: the possible concentration unit measure are:
    - ppb, i.e. part per billion. A unit of ppb corresponds to 1000 mg/L.
    - ug/L, i.e. microgram per liter. A unit of ug/L corresponds to 1 ppb and so to 1000 mg/L.
    - ng/L, i.e. nanogram per liter. A unit of ng/L corresponds to 10^{-6} mg/L.
    - ppm, i.e. part per million. A unit of ppm corresponds to 1 mg/L.

    As can be seen, all the concentration are turned into ”mg/L”.
  - Concentration Type: concentrations based on the active ingredient or formulation, or as the total, un-ionized or dissolved concentration, are identified.
  - Control Type: how the control has been performed
  - Exposure Type: how the chemical is injected
  - Frequency of application in time unit

- Others:

  - Effect: the main effect to consider is the mortality. Other effect to use for RASAR are: ...
  - Endpoint: I focus my work on the lethal concentration (LC50) with the effect Mortality that corresponds univocally to the effective concentration (EC50) with the effect Mortality.
  - Measurement: for each effect ECOTOX database provides a bunch of measurements on the results of tests.

...List of pubchem2d in appendix...

Reporting the concentration to a unique unit (mg/L) allows me to respect what toxicologists want. Indeed, to get the goal of this work I have to set a threshold to figure out the toxicity of chemicals. Usually, toxicologists consider logarithmic threshold like $10^0, 10^1, 10^2, ...$ In any case, I will consider two cases: Binary and Multiclass Classification.

2 Problem Formulation

Most of the interest for ECOTOX database relied on the effect of chemicals in order to kill or to be higly toxic for fishes.

The endpoints I refer to, as previously said, are LC50 and EC50.

Due to the high variance of the data, a regression is not recommended (da vedere e continuare).

Thus, a more reasonable problem may be instantiated thanks to Classification Problem. I do not want to predict the concentration itself but to classify which experiments resulted to have toxic effects on the fishes.