Predicting the exposure concentration of a pollutant on tadpoles based on their behavior and physical traits
Machine learning project

Christophe Reis - University of Lausanne (UNIL)

Abstract

Behavioral and physiological data of 160 tadpoles are used to predict the exposure concentration to which they were subjected. In order to do so, several machine learning tools are used with the objective of developing a reliable model that could increase the sensitivity of the analyses and thus allow improvements in ecotoxicological studies on tadpoles, but also on other organisms. The first results did not allow to arrive at a satisfactory model. However, they are encouraging and offer a perspective of improvement in the future.

key words : machine learning, ecotoxicology, tadpoles.

1. Introduction

In ecotoxicological tests, living species found in the environment are often used to evaluate the toxicity of a xenobiotic (e.g. organic pollutants or heavy metals). These tests consist in exposing individuals to different concentrations and evaluating, over time, its morphological (size, length, etc.) and behavioral (speed of movement, angle of movement, etc.) evolution.
Currently, data analysis only takes into account one of the two aspects at a time. The use of Machine Learning (ML) would have the potential to demonstrate behavioral and morphological changes at the same time. In addition, the use of ML could increase the sensitivity of the data analyses and therefore it would be possible to compare concentrations that are closer to each other compared to what we do today (concentration difference by a factor of 10: 0.001, 0.01, 0.1, 1 mg [mg/L]). In addition, a solvent test is used to determine the solubility of the pesticide.
In addition, the data collection method is based on a software developped by Pennekamp et al. (2015). The Tadpoles are filmed with a video of 60 frames per second and where the difference between two images represents a point. Therefore, the resulting database is quickly very large. ML is again an ideal tool since it is designed to work with a lot of data.

For all these reasons, machine learning is a field that deserves to be studied in order to develop a tool that allows the analysis of morphological and behavioral data according to exposure concentrations.

2. Chloropyrifos and effects on Acetylcholinesterase

The Chloropyrifos is an organophosphor which inhibit the activity of Acetylcholinesterase (AChE) (Giesy & Solomon, 2014), an enzyme that plays an important role in the moderation of neuronal transmission in the nervous system (Trang & Khandhar, 2022). Upon contact with the pesticide, the insects are killed.

Figure 1. Picture of a tadpole.

However, the effect is not limited to the target species, but can also be found elsewhere in the environment and affect other species such as amphibians (see Figure 1). This is why it is important to study the toxicity of pesticides used in agriculture to better understand the problem of its use and its fate

3. Data

The data used in this work are from an experiment conducted by Laurent Boualit, a PhD student at IDYST (UNIL). 160 tadpoles were exposed to Chloropyrifos during a eigth days

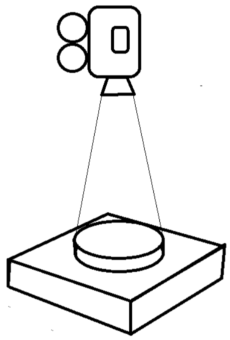

Figure 2. Diagram of the filming.

experiment at five different concentrations (0, 0.0001, 0.001, 0.01, 0.1 [mg/L]). After six, seven and eight days since the first exposure, the tadpoles were placed individually in petri dishes and filmed from above with a light source below (see Figure 2).

Thus, each time the tadpole makes a movement, the camera records the data which are listed in (table 1) and represent five random data. In total, the method allowed to record 269'000 observations with a non-constant number of data per individual. Indeed, some tadpoles were less active than others. Note that the construction of the model is conducted on measurements after six days.

Table 1. Variables collected from the filming program.

Area [pix]	Perimeter [pix]	Maj [pix]	Min [pix]	step_length [pix]	step_speed [pix/sec]	abs_angle [°]	rel_angle [°]	Nom_Conc [mg/L]
87,00	40,63	15,21	7,28	0,63	37,62	3,00	-0,24	0,0001
100,00	44,28	17,96	7,09	2,45	146,76	0,85	-0,20	0,0001
40,00	28,63	10,93	4,66	0,58	35,06	0,75	-2,57	0
77,00	39,11	13,93	7,04	1,25	75,00	3,71	-0,01	0,001
5,00	8,49	3,09	2,06	2,84	24,35	6,27	2,34	0,01

The "Perim" gives the length of the silouhette and "Area" its area. "Major" represents the maximum distance in length and "Minor" in width of the individual. "Step_length" represents the length of displacement and "step_speed" its speed. "abs_angle" and "rel_angle" represent respectively the angle relative to an axis defined on the petri dish and the angle relative to the last position. Finally, "Nom_Conc" indicates the concentration of Chloropyrifos to which the tadpole was exposed.

4. Methodology

4.1. Data preprocessing

After downloading the data, it is necessary to filter the relevant ones. Indeed, the program which films and emits the results returns several variables which are not necessary (e.g. size of the petri dish). It is also necessary to make sure that no value has a "NaN". Since the number of data is huge, all rows that contain NaN are deleted from the database (otherwise, we would transform it into zero to avoid losing to much data). A total of 14,754 lines were deleted (about 5.5% of total data). Furthermore, since there are five different concentrations, the different categories (from zero to five) must be encoded(e.g. 0 = 0, 0.0001 = 1, 0.001 = 2, and so on.).

Once the data processing is finished, we must now choose the criteria that will be used to predict the concentration and divide this data set (also taking the concentrations that are related) into three sets: Train (70%), Test (15%) and Validation (15%).

4.2. Model-based approaches

The first method that is used in this work is a simple classifier using a few classification algorithm such as RandomForestClassifier or DecisionThreesClassifier. The model will be first trained in a train dataset before being applyied to a test dataset. A confusion matrix is generated and the accuracy score is calculated. According to the score, a search for better hyperparameters (HP) will be apply to find a the best ones. Once the accuracy seems acceptable (at least 80%), the model will work on the validation dataset (unknown by the model) to see if the accuracy is still good. Finally, the model will be tested on the two others days (seven and eight) which are two datasets that the model never saw.

In order to take the problem from several angles, others methods will also be tested. it would be interesting to explore the possibility offered by the softmax regression which is a kind of logistic regression but for multiclass (more than two). The two approach could then be compared.

Table 2. Accuracy result with RandomForestClassifier.

	Variables name	RTC result [%]
Behavior features	Step_speed	39,8
	Step_length	28,4
	Abs_angle	25,1
	Rel_angle	25,6
Morphological features	Perimeter	28,2
	Area	27,1
	Major	35,5
	Minor	31,4
Combinations	Step_speed + Major	76,4
	Step_length + Minor	74
	Step_speed + rel_angle + Major	75,6
	Step_speed + Major + Minor + step_length	77,3

5. Results

5.1. Ensemble learning (classification)

First, a classification was run using the RandomForestClassifier for each of the variables independently. Then, some multi-variables RFC are processed to check the accuracy. All the results are summarized in (table 2). From the different results, it is possible to say that having two or four variables doesn't change the accuracy so much (we stay around 75-77%). It is therefore that the two highest values in each category are retained for the rest of this work. For the first tests, RFC is used. After the first run through the train and test dataset, we reach an accuracy of 77.5% for the test set. We can therefore conclude that the model is satisfactory but does not have sufficient accuracy to accept it. To improve the model, a grid search CV is conducted to find the best hyperparameters around the default one since the accuracy is already good. Unfortunately, the running time was much too long, even using all the eight processor, and no results came out of this HP search.
Therefore, another classifier has been adopted : The DecisionTreesClassifier. This classifier gave an accuracy of 76.3% with default hyperparameters. Which is slightly over the RFC accuracy. Again, a grid search CV is conducted to find the best hyperparameters around the default one. However, as for the RFC, the hyperparameters found were never able to make the model evolve and never exceeded 30% accuracy (even by doing a RandomizedSearchCV).

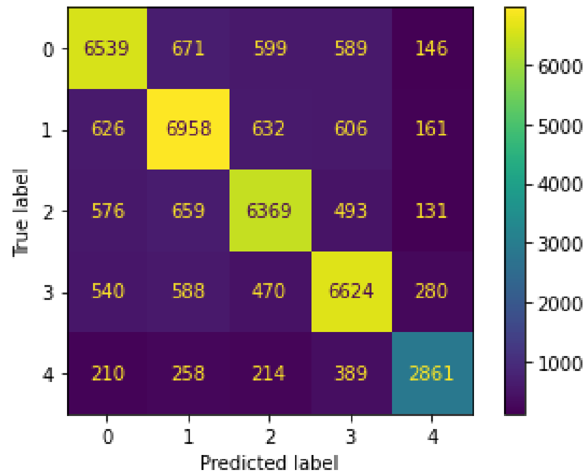

Figure 3. Confusion matrix using RFC on validation dataset.

The Figure 3 shows the confusion matrix obtain with the best model (RandomForestClassifier with default HP). The label 0 is equal to the concentration of 0 [mg/l], the label 1 is equal to 10^{-3} [mg/l] and so on until 4 that is equal to 10^{-1} [mg/l]. We can see that the model confuses all concentrations with the same probability with a few exceptions. Indeed, if we take the last line, we see that the model predicts wrongly the concentrations from 0 to 10^{-2} [mg/l] the same number of times. We can see that despite a factor of 10 between the concentrations, it is still difficult to see a difference when taking into account the behavior and morphology.

5.2. Softmax regression

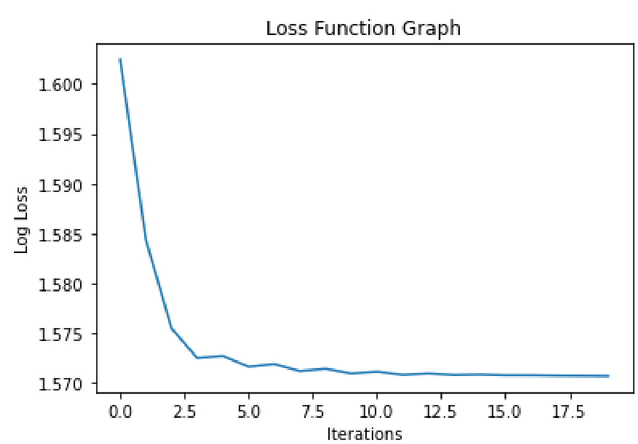

Figure 4. log_loss function for the SMR.

Unfortunately, the use of the softmax regression model did not allow to obtain a more conclusive result. Indeed, even by adapting the learning rate and the number of epochs, the precision of the model never exceeds 24% of accuracy. Moreover, (Figure 4) shows us that the log_loss of the model starts from a value that is far too high, but also that the attenuation of the latter over the epochs is far too weak to offer an ideal model.

It may be that the model does not quite fit the dataset to be analyzed. Further investigations should be conducted to improve the softmax regression model in order to use it. In any case, this is a possibility that remains to be explored in the future.

5.3. Application to data from day 7 and 8

Now that all possibilities are exhausted, the best model obtained so far is kept to be applied to the next days data. This is to demonstrate that the model is well trained. Therefore, the best model obtained is the classification using RandomForest with default hyperparameters with an accuracy of 77.5% using the four variables in the last row of Table 2.
From then on, the model is again trained on the train set of say six. Then it is applied to the entire data of day seven and eight without training. The accuracy is the same for the two run (77.5%). We can therefore say that the model using RandomForest offers a constant model that gives an equal result between the data of day six on which he trained and the days seven and eight that he never saw.

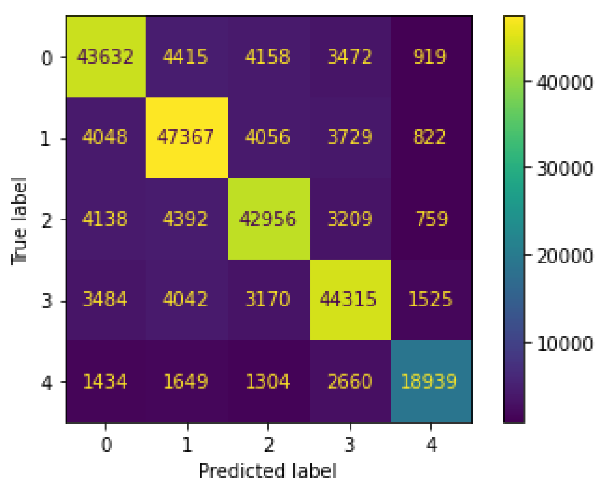

Figure 5. Confusion matrix using RFC on day 8 dataset.

To have a visualization of the errors, Figure 5 shows the confusion matrix of the model. We see that the error matrix for eight days is very similar to the one for six days on the training dataset except for the number of predictions since the model is applied to the whole database. We can therefore conclude that if the model is improved in the future, the result on days seven and eight will also be improved in the same way. This is very promising.

6. Discussion

Through this project, several aspects of machine learning can be highlighted. First of all, we could see that the RadomForest is an algorithm that works very well in a natural way without the need to modify it. This shows its strength and explains why it is one of the most robust models known. However, other classifiers such as the DecisionTrees are still models that stand up well.
Despite the good basic accuracies, the hyperparameters could not be improved using GridSearchCv or RandomizedSearchCV. This may be due to poor initialization of the HP to be searched. For the next research, it would be interesting to continue the research in order to obtain a precision higher than 80% (even close to 90%).
In a second time, the research of another method allowed to develop a classification with several labels: the softmax regression. However, once again, the result is not up to standard. A lot of work remains to be done in order to make the model truly optimal. It is again an approach that deserves to be studied and optimized.
Also the results obtained in section 5.3 are very encouraging. Indeed, the model works well on data that it has never seen before. With an improvement of the different models seen in this project, it would be possible to develop a powerful model that would work on all databases.
Finally, despite the multiple angles of attack already taken, there is still one method that has not been explored: the neuronal network. This model is very efficient and can largely surpass those seen in this project. Unfortunately, it could not be treated here. In future research, after having continued to develop the first two approaches, it would be wise to try it in order to do a comparison.

7. Conclusion

To conclude, we can say that machine learning is definitely a very powerful and useful tool that can be adapted to a large number of study cases. However, it requires a lot of resources and time if we want to have a model that is not only operational, but also optimized. Indeed, even if the results obtained for the prediction of days seven and eight after a training on day six are satisfactory, there is still a lot of optimization work to be done in order to get a very satisfactory result.
Finally, the number of possible approaches offered by machine learning allows the user to try several methods and thus get several results as well as a freedom of creativity that has few limits. And, in the absence of convincing results from the trials carried out in this project, this is certainly the best conclusion that is possible here.

8. Python code link

Link to the GitHub page

9. Acknowledgements

First of all, I would like to thank Laurent Boualit for his enthusiasm since the moment I proposed this project. The results are not satisfactory for the moment, but I promise, the work continues!
I would also like to thank my peers for their precious feedback. They helped me to improve not only the final report that you have read, but also the machine learning part.
In addition, I would also like to thank Tom Beucler and Milton Gomez Delgadillo for their responsiveness and availability throughout the semester to answer the thousands of questions and spend time trying to improve my models.
Finally, I would like to thank three people who will probably never hear about me and this project, but who helped me a lot for the ”Softmax regression” part with their online articles: Suraj Verma, Lily Chen and Arthur Juliani.

References

Giesy, J. P., & Solomon, K. R. (Eds.). (2014). *Ecological risk assessment for chlorpyrifos in terrestrial and aquatic systems in the united states* (Vol. 231). Springer International Publishing. https://doi.org/10.1007/978-3-319-03865-0

Pennekamp, F., Schtickzelle, N., & Petchey, O. (2015, June 4). *BEMOVI, software for extracting behavior and morphology from videos, illustrated with analyses of microbes*. Retrieved May 19, 2022, from https://onlinelibrary.wiley.com/doi/full/10.1002/ece3.1529

Trang, A., & Khandhar, P. B. (2022). Physiology, acetylcholinesterase. *StatPearls*. StatPearls Publishing. Retrieved May 11, 2022, from http://www.ncbi.nlm.nih.gov/books/NBK539735/