https://codalab.lri.fr/competitions/623

TEST REPORT OF GAIASAVERS CHALLENGE BY XPORTERS

Plankton Classification Challenge

[Florian Bertelli][1], [NGaspard Donada–Vidal][2], [Ghassen Chaabane][3], [Moez Ezzeddine][4], [Ziheng Li][5] and [Gérémy Hutin][6]

ABSTRACT

In the following document, our aim is to test the challenge made by the team Gayasavers, as part of the Master 2 AIC in University Paris-Saclay. This challenge is a classification challenge on a large image database of various kind of plankton. The website of the competition is https://codalab.lri.fr/competitions/623

STARTING KIT AND SCORING PROGRAM

- The Jupyter Notebook works fine and produces sample submission. The sample submissions work and produce the same result as in running in local. The metric is the same in both methods
- If we compare this with a random methods which output random label ie a arbitrary label chosen in between 7 different labels, the result should be about 0 (with a balanced accuracy score), and here we got ith the basic model 0.54. Thus our results is better than a random one.
- The jupyter notebook provides us basics knowledge about the features extracted from the images like the sum according to some axis. Furthermore it is well explained what the features are and how they are represented. So the iPython notebook provides data preprocessing leading to a feature-based representation, which are understable.
- It will be appreciable to have a few other images to see different images corresponding each to different labels. It will be great to plot the features according to the corresponding target to see whether they have influence on the target. Maybe plot a correlation matrix should be interesting too, to determine which parameters are correlated.
- All the code is present in the following github : https://github.com/Tahkare/Projet-AIC

DATA COMPETITION AND MAKE BUNDLE

- The dataset is in the correct AutoML format which allows it to be read through the Jupyter notebook with the data-io package.
- The competition bundle is on github.
- The dataset for this challenge does not contain any missing data so there are no problems associated to this.
- We have tried to upload on the CodaLab the basic starting kit results. We get this results.

17	Oct. 8, 2019, 8:30 a.m.	xporters	11232	sample_result_submission_19-10-08-10-29.zip	Finished	False	0.5463
18	Oct. 8, 2019, 8:37 a.m.	xporters	11234	sample_code_submission_19-10-08-10-29.zip	Finished	True	0.5463

Everything about the codalab compeititon is working well.

- Then we're interested about data leakage. Data Leakage happens when for some reason, your model learns from data that wouldn't (or shouldn't) be available in real-world scenario. It often happens when you are using time related series. Here, we are just using basic images, which are available in real life (we don't use any other informations initially to build the dataset). Furthermore, if we consider the results, it doesn't appear that there are too perfect. Indeed, with the provided model in the starting kit, we get a score of 0.54, so data leakage seems impossible.

WORKFLOW AND DOCUMENTATION

- The overview and the description are well done, it is now clear that we are doing a classification problem on the various forms of plankton.
- The challenge has two phases, one of development and one of test as wanted. However the date seems to be wrong.
- The part evaluation has clarity, self-sufficiency, and soundness. Maybe one should add precision about the method that will be used to score the model (precise which kind of accuracy score it is).
- The \overview" page contains credits (1) the team member names, (2) a contact email, (3) credits to the database donors + URL.
- The page 'Terms and conditions' is there and unchanged

CONCLUSION

The planton classification challenge created by the team GaiaSavers appears to be well built, and very interesting. We strongly recommend to do it.