ZOOPROCESS / PLANKTON IDENTIFIER PROTOCOL
For
COMPUTER ASSISTED ZOOPLANKTON SORTING

Garcia-Comas and Picheral
2013/03/28

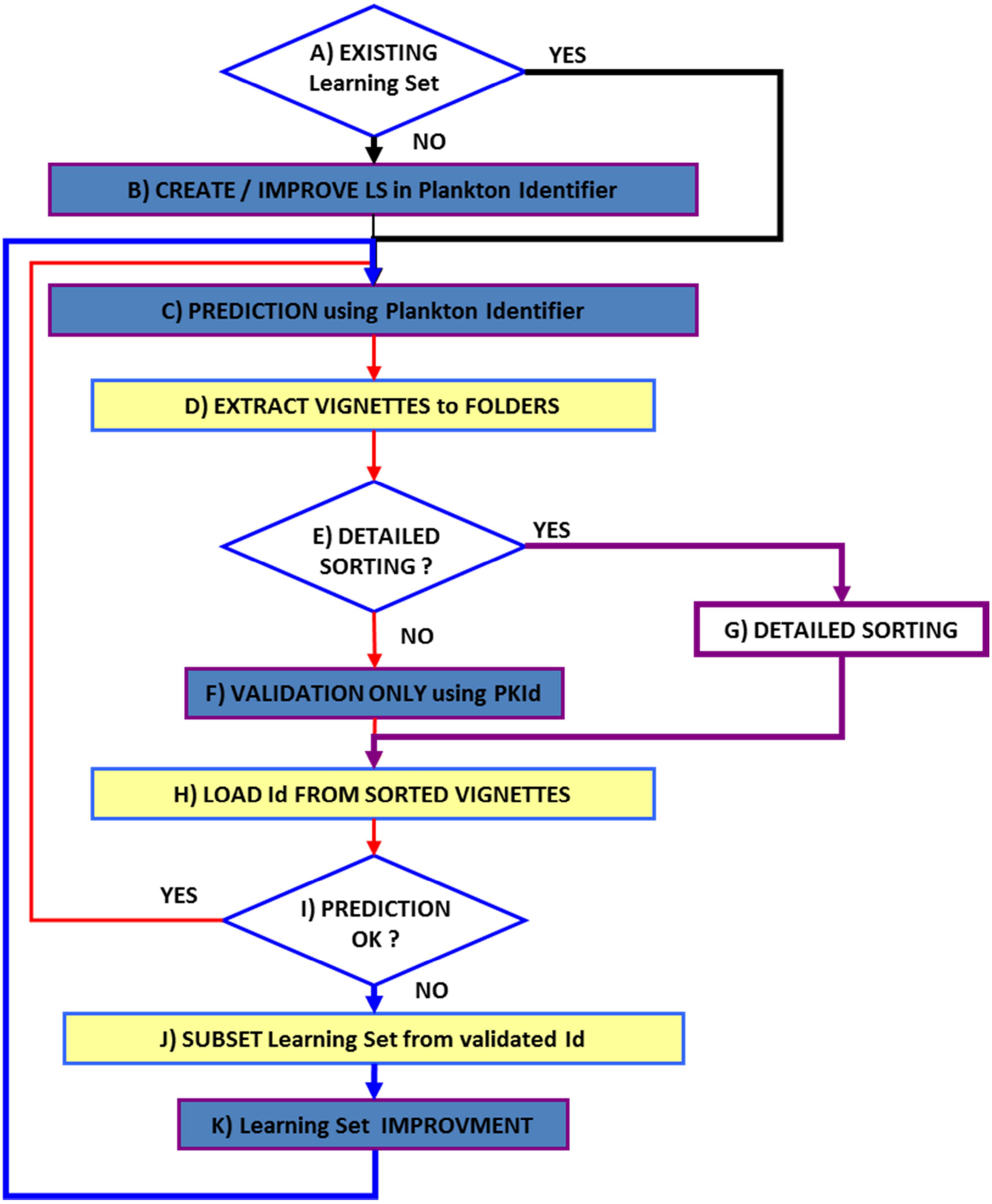
We describe below the general strategy to perform computer assisted sorting of plankton using the Zooprocess/PKId tools. This method is applicable for all sorts of images which can be acquired with diverse instruments (ZooSCAN, FlowCAM, UVP5, ISIIS, microscope) and processed with Zooprocess.

While with automatic recognition we try to recognize very fast a few well distinguished groups, when we validate data we can add new groups or separate vignettes of a group in more specific ones (e.g., copepod transformed in three groups: cop_calanoid, cop_oithona, cop_oncaea).

READ : ***J. Plankton Res.*** Gorsky et al. 32 (3): 285.

Content

1 GENERAL DIAGRAM


A) EXISTING Learning Set
YES
NO
B) CREATE / IMPROVE LS in Plankton Identifier
C) PREDICTION using Plankton Identifier
D) EXTRACT VIGNETTES to FOLDERS
E) DETAILED SORTING ?
YES
NO
G) DETAILED SORTING
F) VALIDATION ONLY using PKId
H) LOAD Id FROM SORTED VIGNETTES
YES
I) PREDICTION OK ?
NO
J) SUBSET Learning Set from validated Id
K) Learning Set IMPROVMENT

A) The first step is to check if you can utilize an existing Learning Set (LS) that could run on the samples you have scanned. A coarse LS can be utilized to perform an initial prediction of some samples. The resulting Identifications (Ids) on these samples will then be checked and utilized to create a new, more accurate, LS. Such coarse LS can be downloaded from the ZooSCAN website.

B) If you consider that you do not have any usable LS, you will create this initial basic LS containing at least 50 objects per category and a limited number of categories (~10). This operation should take you less than half a working day including Cross Validation tests at some stages.

C) Predict the Ids for some samples (about 5, no more) that you have scanned. These samples should be as much diverse as possible (e.g., different location and/or dates) to cover most of the community variability that you may have in your dataset.

D) Extract vignettes to folders according to the prediction you have done in order to visualize the prediction.

E) You can decide whether you add categories (detailed sorting) to the ones you predicted, or you just limit the validation to the predicted ones.

F) Validate the predicted Ids sorting the vignettes using PKId or XNVIEW image browser.

G) Same as in F.

H) Load Ids from the vignettes you have sorted (adding Ids to measurements in the rows of the PID file)

I) According to the results (% of Ids corrected) and your personal estimation of the validation workload, you may consider the LS to be satisfactory or continue improving it. In the first case, go back to step C) to predict other set of samples that you will then validate.

J) If you are not happy with the prediction (initial LS or LS from other project), you will try to improve it. A good method is to utilize the vignettes you have sorted and create a LS from them using the dedicated Zooprocess tool.

K) Improve the "automatic" LS that has been created by Zooprocess to keep only "good" items. Perform some Cross Validation to evaluate again the new LS as for the initial one.

2 STEP A & B : CREATE AN INITIAL LEARNING SET

We strongly recommend to spend very short time (half a day) on this task, and rather to validate all your samples after automatic recognition (prediction). In our procedure, automatic recognition is a mean to sort organisms faster. We call it "COMPUTER ASSISTED ZOOPLANKTON SORTING" as the human is who finally decides the classification.

You need to switch to "ADVANCED mode" for your project to be able to use the related Zooprocess tool (do not forget to switch back to "USER mode" afterwards).

1. Copy, if not done yet, the "ZooSCAN_liste_ident.txt" file from the config folder of your project into the "program_files\Plancton_Identifier\liste" folder.

2. Enter your project folder and go to "Pid_process" and into "Pid_results". COPY the .pid files of the samples that you want to use to build your Learning set (a small representative subset of your whole set, e.g., seasons represented) and paste them in "Unsorted_vignettes_pid".

3. Start Zooprocess if necessary. Choose the menu "Extract vignettes for Plankton Identifier" (ADVANCED mode only).

4. Keep default options. Change Resolution to the one you have used for scanning if necessary. You can increase a bit your Gamma if you think the contrast of the vignettes is not enough (e.g., 1.2), but we usually leave it at 1.1 par default. If you don't want to use all the objects (e.g, you prefer to create a LS with subsets of many samples instead of all the objects of a few samples).

5. Your vignettes are now with the related .pid files in “Unsorted_vignettes_pid” directory. It is recommended to store all the extracted vignettes in a new folder inside “Unsorted_vignettes_pid” and name it as “unsorted_YYYYMMDD” to keep track of it. Important! move also the associated .pids to the new folder.

6. Create a subfolder in the “Learning_set” folder named “YYYYMMDD_HHMM_initial” to keep organized too.

7. To CREATE the Learning Set in Plankton Identifier (PKId), close Image J and open PKId (See PKId manual on its website).
   a. Click on LEARNING and select the folder just created.
   b. Click on the folders icon up on the right to create groups (i.e., categories as copepoda, cladocera.. etc). We recommend using the default naming list which is created in the configuration folder of the project and you copied into PKID list folder.
   c. On the left (i.e., Unsorted thumbs) select your folder in “unsorted vignettes” (i.e., unsorted_YYYYMMDD). Select vignettes (no more than 50 each time because they are copied and not moved and thus you could have windows memory problems) and move them to the group they belong to.
   d. Once you have finished sorting ALL the vignettes, click on the icon on the right bottom, “create learning file”. We recommend naming it as Learningset_YYYYMMDD_HHMM to keep track of it. Save it in the folder created above (note that this file is a concatenation of all the variables from the PID files of vignettes used to build it, and that a last column has been added with the Id of each object). You will have a pop-up wndow: “JOB COMPLETED, continue sorting”,click “NO”.

8. To EVALUATE a LEARNING SET before using it to predict Ids of organisms
   a. Click on DATA ANALYSIS.
   b. On the learning set box, you select as Learning file your LS.
   c. Select the method Cross-validation4 (Rndm tree) (on the left bottom of the main box). The method consists of one part (random) of the learningset recognizing another (2 folds), and this is repeated 5 times to obtain robust statistical values.
   d. Click on “Start Analysis”. Include in the name of the analysis the date to keep track. Select to SAVE RESULTS IN “Prediction” folder of PID_process folder of your project.
   e. After analysis is completed, quit data analysis.

9. Click on SHOW REPORT to visualize the results of the Cross Validation.
   a. Select the Analysis_name.txt you have created. Your web browser will open.
   b. Click on cross-validation. You can see true classification (rows) versus automatic classification (columns). The recall is the % of organisms belonging to a group that were automatically well recognised, whereas the 1-precision is the % of organisms classified by the algorithm as a group that is not so (contamination in a group). **Recall should be above 70% for all groups, and contamination (1-Precision) below 20% to keep a category.**

3 STEP C : PREDICTION of IDENTIFICATION

Using the LS we created or recovered from another dataset we predict a bunch of samples of the project.

1. Click on DATA ANALYSIS.
    a. Select on the left top box the LS folder and the selected LS.
    b. Below, you select in "Pid_results" folder the samples PID files that you want to predict (recommended maximum of 20-30 pid files at a time to avoid memory problems).
    c. Check the variables that will be used to classify the objects (untick position variables!).
    d. Select the SpvLearning4 (Random forest) method.
    e. TICK "save detailed results for each sample". A file will be created for each sample. This *dat1.txt file will be the metadata concatenated to the table of objects in the .pid file with a last added column containing the automatic classification of each object (row). Save results, as Analysis_YYYYMMDD_HHMM to keep track of the date, in "Prediction" folder of Pid_process of your project.
    f. When PkId has finished you can close it.

2. Move the PID files already predicted into the "PID_predicted" folder to avoid predicting them again!

4 STEP D to I : VALIDATION of IDENTIFICATIONS

Once the prediction is made, it has to be validated by a person capable to check Ids. At this step, the expert corrects missorted vignettes and can also add new IDs if desired.

1. Copy the Analysis_sample_dat1.txt files of the samples to be validated from the "Prediction" folder to the "Pid_results" directory.

2. Run the Zooprocess tool: "extract vignettes in folders according to prediction". Leave the default settings but check resolution. Press OK. The folder "sample_YYYYMMDD_HHMM_to_validate" with the objects predicted is created in the "Pid_process\ sorted_vignettes" folder. The predicted sample_dat1.txt files have also been automatically copied to Dat1_extracted in "Pid_results". By default, empty folders are also created using the IDs list from the config folder.

3. Delete the Analysis_sample_dat1.txt files from the "PID_results" folder. (Delete all for ZooSCAN, and only those you extracted for other instruments).

4. To accelerate validation of samples, we recommend using XnView (free software) instead of copying vignettes from one folder to another in Windows Explorer. It moves the vignettes and thus, you can move many more vignettes in one selection without memory problems. Open XnView and select the folder that you want to validate (i.e., sample_YYYYMMDD_HHMM_to_validate). Go through each of the subfolders (category) and check the vignettes. If classification is wrong or can be improved, move the vignette to the right folder. When done, change folder name from "*_to_validate" to "*_validated".

5. Once you have finished, close XnView and go to Zooprocess. Select "Load identifications from sorted vignettes". Select the sample you want to process (or ALL

folders). Zooprocess will tell you the percentage of vignettes that you resorted during validation (including misclassified vignettes and vignettes resorted in new finer groups defined by the expert). This information can help you deciding to improve the LS or not. A 20-30% error rate is common for diverse datasets (>25 predicted categories)

6. The sample_dat1.txt files have been renamed by deleting the "analysis_" at the beginning of the name (final data name) and copied in the Dat1_validated of "Pid_results". These final tables have a new last column which contains the expert classification of each object.

5 STEP J to K : CREATE A NEW LS FROM VALIDATED DATA

If you consider that you will save time in future validation tasks on the project by creating a new LS, we describe below a simple procedure to accelerate this operation. This operation should not last more than two hours.

Switch to ADVANCED mode (don't forget to switch back to USER mode once finished).

1. The Zooprocess tool: "CREATE subset of a Learning Set from identified vignettes (random)" will randomly subsample vignettes from validated samples (and renamed "validated") and place them in a new subfolder of the "LEARNING_SET" directory named "YYYYMMDD_HHMM_random_N" (N being the chosen number of vignettes). Consider ending with no more than ~ 300 vignettes for each category in your final LS.
2. Open this LS in PKId to ("clean the new learning set"):
   a. Remove folders containing very low numbers of organisms.
   b. Remove vignettes that are badly identified.
   c. Remove vignettes that are not representative of what should be predicted.
3. Test the new LS using cross-validation as described above.

As you moved some vignettes and probably deleted others, your LS will no longer be fully balanced. If you want to homogenise this new LS with the same number of organisms in each category, just run again the same Zooprocess tool selecting now the random LS that you just created instead of the default "Sorted_vignettes" folder.