Common project - AI & Optimization – Data Analysis

Isep Paris, 28th April 2023

Automatic ladybug classification and spot counting

Deadline : Monday, June 5th 1 p.m.
Work alone, or in teams of maximum 3 students

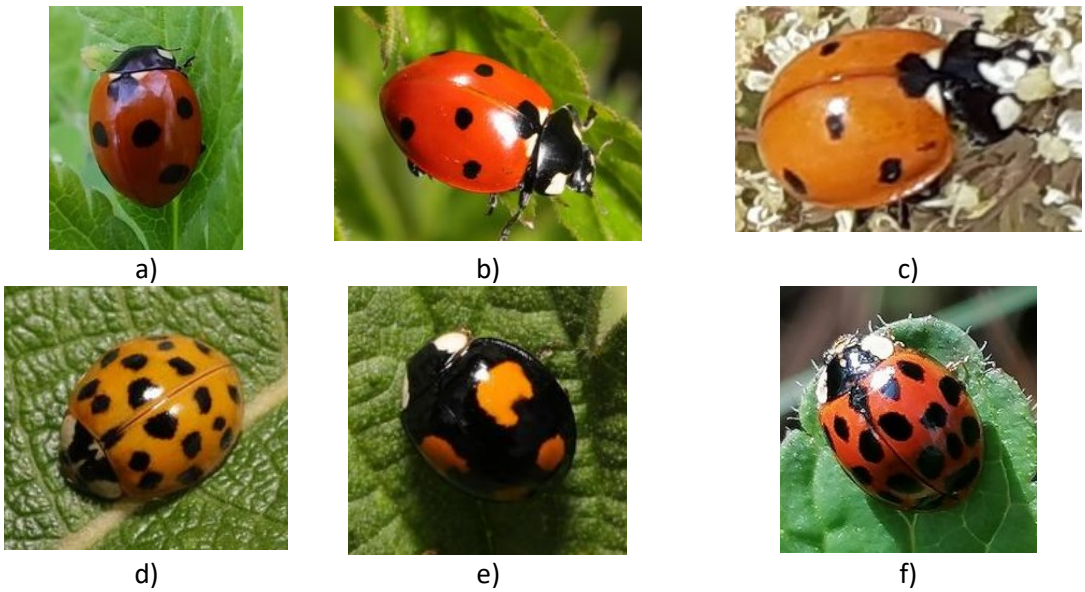
Project sessions: - Wednesday, May 17th
- Monday, May 22nd,
- Wednesday, May 24th,
- Monday, June 5th

I. Context

Harmonia axyridis (Figure 1) is a ladybug specy that was massively introduced in Europe in the 80s' for aphid biological control. Indeed, adults and larvas are formidable predators, eating a huge amount of insects considered as harmful for human agriculture, but also eggs of local species such as Coccinella septempunctata (Figure 1). Due to their voracity, high fertility rate and behavior, Harmonia axyridis ladubugs are now considered as an invasive specy in Europe – competing with local species for food and space [1].

Thus, studying the ladybug population is crucial to monitor biodiversity. This process can be automatized from pictures – in fact, ladybugs species are visually easy to identify. For example, Coccinella septempunctata ladybugs share a common pattern – red elytra with seven black spots. On the contrary, Harmonia axyridis ladybugs may differ in elytron colors, and spot shapes, number or colors. Thus, spots can be used as a relevant feature to classify ladybugs.

In this context, the goal of this lab is to develop a Machine learning approach to automatically detect the specy and count the number of spots on an ladybug image.



a) b) c)

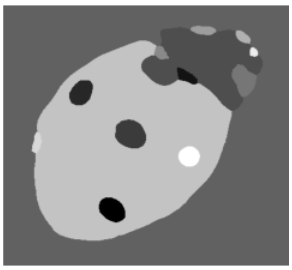
d) e) f)

Figure 1: RGB images of Coccinella septempunctata (a,b,c) and Harmonia axyridis (d, e, f) ladybugs

II. Material

The dataset, available on the Moodle pages of the courses II.2413 Data Analysis and IG.2411 AI & Optimization, is made of RGB images (im-*bug id*_im.png) and segmentations (im-*bug id*_seg.png) of 200 ladybugs. For each ladybug, the segmentation is an image of same dimension as the RGB image where each ladybug region of homogeneous color is associated with a distinct number (Figure 2).

(a)



(b)

Figure 2: RGB image of a ladubyg (a) and its associated segmentation (b).

III. Training phase

The project is divided into two steps: training and test. In fact, it can be seen as a mini-challenge! The training phase takes place during the first three project sessions.
The training set is made of images and segmentations of 200 ladybugs. They are available in the « training_data » folder. The specy associated with each ladybug_ Harmonia axyridis or Coccinella septempunctata_ is given in the « training_labels.csv » file.
Your work during the training phase may be divided into up to five parts:

a) Spot counting on the training dataset
   Visually compute the number of spots on each training image, and complete the «spot_number » column of the « training_labels.csv » file.

b) Data exploration (not mandatory but recommanded)
   Get familiar with the dataset. You can for example create a « quality control » folder, where the images are split into 2 subfolders according to their associated species. It helps you visualizing the difficulty of the task, and choosing relevant features. You can also choose to discard some images that look irrelevant for training… but be careful, similar examples could appear in the future test set!
   You can also answer this useful question during this part: what is the label distribution among the training dataset? It will be roughly the same among the test set.

c) Feature extraction and visualization (**mandatory**)
   - Extract **homemade** features from each training image. Thus, **you have to implement your own features!** Of course, you can choose features implemented in the previous lab and use very basic libraries

such as numpy.
When all the features are collected for all the available images, you are strongly recommanded to save them in a csv file.

- **Vizualize** your training data using the feature you just created with methods from the data analysis class. Use these visualizations to infer the difficulty of the clustering and classification task from part d).

d) Training of a machine learning algorithm (**mandatory**) -
Train machine learning algorithm(s) on the selected features to
   - classify the associated images into Harmonia axyridis or Coccinella septempunctata
   - Propose a clustering of your data
   - count the number of spots.

The selected algorithm(s) should have been studied during the Optimization & AI or Data Analysis course.

More especially, you have to be careful about the following points:
- feature selection, transformation or normalization (if needed)
- hyperparameter tuning
- algorithm selection
- optimization algorithm
- cross-validation

e) Test on the future dataset (not mandatory but recommanded)
Prepare the code to apply the trained model on the test dataset. It has to automatically save the results in a « test_labels.csv » file (same column names as « training_labels.csv »).

IV. Test phase

The test phase takes place during the last 30 minutes on the last project session, namely on Monday, June 5th.
A folder containing test images and associated segmentations («test_data ») of 50 unseen ladybugs will be made available on Moodle. Apply your trained model on these images, and save the predicted labels in a « test_labels.csv ». The structure of this file should be similar to the one of the « training_labels.csv » file, namely with a first column «Image_id» containing the identifiers of the test images (arranged alphabetically), and a second column «Label» with the predicted label for each test image. **Penalties will be applied if the structure of the « test_labels .csv » file is not respected.**

V. Expected outputs

At the end of the last project lab session, one single member of your team has to depose two files on Moodle.

a) **The ipynb code file**
A .**ipynb** file containing the code for the **training phase** (**feature extraction** and **ML training**), and **test phase**. This code has to be well structured and commented. More precisely, add in markdown cells any comment useful to justify your choices.
You have to clearly mention the names of all the members of your team in the code file.
As usually, assess that the following requirements are fulfilled:
- Run your Jupyter notebook before sending it back (kernel → Restart & Run all).
- Check that there is no error message.
- Do not write any space or special character in the name of your file
- Ensure that each line of code does is not (roughly) composed of more than 80 characters

b) **The csv prediction file**
A csv file containing three columns, the first one for the test image identifiers, the second one for the predicted labels (Harmonia axyridis or Coccinella septempunctata) and the third one for the predicted number of spots. The structure is the same as in the training_labels.csv file.

VI. Grading

The grading will be divided into three parts:
- 7 points: justification of the choices made for feature and machine learning frameworks, compliance with the instructions, quality and clarity of the ipynb file.
- 3 points for the clustering and visualization tasks
- 5 points - classification task: relative accuracy obtained for your group i on the test set compared with the maximum relative accuracy obtained among all the groups:

$$5 * Acc_i / \max_{j \in \text{nb groups}} (Acc_j)$$

- 5 points – spot counting task: relative Root Mean Squared Error (RMSE) for your group i on the test set compared with the maximum RMSE obtained among all the groups:

$$5 * RMSE_i / \max_{j \in \text{nb groups}} (RMSE_j)$$

VII. Reference

[1] https://france3-regions.francetvinfo.fr/grand-est/faut-il-se-mefier-coccinelles-asiatiques-1652776.html