

Plankton Identifier

Presentation

Plankton Identifier is a free software which allows the automatic identification of objects (plankton and/or particles) from a set of images with their associated data. Therefore, before you can use *Plankton Identifier*, you must extract object data (PID format) and object thumbnails (JPEG format) with a dedicated hardware-software combination such as ZooScan and ZooProcess (<http://www.ZooScan.com>).

Once data acquisition is done, *Plankton Identifier* helps the user to build a Learning File (set of objects identified by an expert) through a safe interface which prevents from data loss, erroneous duplicates etc. Then, using the learning file, *Plankton Identifier* allows the automatic recognition of an unlimited number of objects from one or several data files. Additionally, the user can optimize the recognition by many ways (by choice of different data analysis methods, selection of variables, variable transformation, way of grouping objects etc.). Results are stored in one or several text files which can easily be imported in other softwares (for graphics, time series analysis etc...). A detailed statistical report is also generated (html format allowing copy-paste).

NOTE: Although Plankton Identifier has been initially developed to identify Planktonic organisms, it can virtually identify any kind of objects.

Supervised learning algorithms are not implemented in *Plankton Identifier* itself but are those of the free data mining software TANAGRA which is used in batch mode. This is the reason why you must install Tanagra 1.4.12 or above on your computer before using *Plankton Identifier*.

NOTE: Experimented users can define their own data analysis methods for Plankton Identifier using Tanagra tdm format.

Downloads

Version currently available is 1.2.6. It is a Win32 application running under Windows XP, Windows 2000 and Windows 98. A screen resolution of at least 1024x768 (96 ppi only) is highly recommended. There is no UNIX, LINUX or MacOS version.

To install *Plankton Identifier*, download the installer on your computer. Then, execute the installer and follow the instructions.

Seven recognition methods (all tested) are currently installed with *Plankton Identifier* as well as several methods to test classifier performances. Some demo files (zooplankton JPEG images and PID files) are included and will be installed in a 'Demo' subfolder.

An additional application, *PID Viewer*, can be downloaded separately. *PID viewer* helps to visualize PID files information (header, data, and basic statistics). Once *PID viewer* is installed on your computer, you just have to double-click on PID files (Sample or Learning) to easily explore and/or compare their contents, even from *Plankton Identifier*.

NOTE: PID Viewer is not an editor. Manual edition of a PID file is not recommended but if necessary you can use any usual text editor.

Sources

Plankton Identifier has been developed using Delphi 2005 PE (Codegear from Borland), which is free under conditions, and the free components Virtual Treeview (Soft-Gems) and VirtualShellTools (Mustangpeak). *PID viewer* uses the free Delphi component Toneinstance (home.com). The installers have been created using Inno setup 5.1.9 (Jordan Russell's software). *Plankton Identifier* and *PID viewer* source codes are available on demand for non commercial purpose.

Please, don't forget to mention Plankton Identifier when you produce results, notably in scientific publications. You can refer to Plankton Identifier as follow:

Gasparini Stéphane (2007). PLANKTON IDENTIFIER: a software for automatic recognition of planktonic organisms. http://www.obs-vlfr.fr/~gaspari/Plankton_Identifier/index.php

Acknowledgements

Thanks are due to Elvire Antajan, Ricco Rakotomalala, Marc Picheral and Gaby Gorsky for support and contribution to this project.

Stéphane Gasparini
Laboratoire d'Océanographie de Villefranche
UMR 7093 - CNRS / Université Pierre et Marie Curie - Paris 6
BP 28 - 06234 Villefranche sur mer Cedex - France

USER GUIDE

http://www.obs-vlfr.fr/~gaspari/Plankton_Identifier/userguide.html

Stéphane Gasparini and Elvire Antajan

January 2008

MAIN WINDOW	4
LEARNING	6
Select Folder Window	6
Learning Window	7
Unsorted Thumbnails	7
Groups (subfolders) creation	8
Thumbnails sorting	9
Modify the Learning set	10
Cancel action	10
Create Learning file	10
DATA ANALYSIS	11
Select learning file	11
Select sample file(s)	11
Select a method	12
Original variables	13
Customized variables	13
Identification Groups	14
Launch analysis	15
SHOW REPORT	16
APPENDIX	18
File formats and file name	18
PID files	18
Thumbnails	18
Learning Files	19
Folders organization	19
Recommendations	19
Learning Set	20
Build customized analysis method (tdm)	20
FAQ	22

MAIN WINDOW

When launching the application, 3 boxes are available: "Learning", "Data Analysis" and "Show Report" (Fig.1). These 3 steps can be run independently. However, "Data analysis" uses files generated by "Learning", and "Show report" uses files generated by "Data Analysis". Thus, the very first time you run *Plankton Identifier* you must begin with "Learning".



Figure 1: Main Window

A menu is also available:

Program > Settings allows to define the location of Tanagra.exe as well as a default folder for thumbnails, PID files and results.

- **Tanagra Path:** If more than one version of Tanagra is installed on the computer, you can select the version you want to use. Click **Browse**, browse the hard drives until you find Tanagra.exe, select it and click **OK**.

NOTE: If you did not install Tanagra in \Program Files\Tanagra, or if you did not install Tanagra at all, this window will pop up automatically when running Plankton Identifier (Fig.2).



Figure 2: Settings Window

- **Default folder:** The default folder will be the starting folder for the different steps the very first time you use them. When installing *Plankton Identifier*, the default folder is the "Demo" folder of the *Plankton Identifier* install directory. Thereafter, the starting folder will be the last folder successfully used except if you have deleted that folder in between (then it will be the default folder). If none of these folders exists anymore, "My documents" will become the default folder. To change the default folder, click **Browse**, browse the hard drives until you find the folder you want as default folder, select it and click **OK**.

Program > Exit closes *Plankton Identifier*.

Tools > Import Tanagra Data Mining diagram helps to import data analysis methods created with Tanagra and saved as a tdm into *Plankton Identifier*. When the window opens, browse the hard drives, select the tdm file to be imported, and click **open**. The routine verifies if the selected tdm can be used in *Plankton Identifier*, adapts it if necessary, and makes a copy in the appropriate subfolder (*Plankton identifier* install directory \ Classifiers).

NOTE: A successful import does not mean that the method will work. Please refer to the section "Build customized data analysis method" for more information.

Tools > Import Name List helps to import a name list (for groups / subfolder creation) into *Plankton Identifier*. When the window opens, browse the hard drives, select the text file to be imported, and click **open**. The routine verifies if the selected text file can be used in *Plankton Identifier*, and makes a copy in the appropriate subfolder (*Plankton identifier* install directory \ Lists).

NOTE: The text file must contain group names separated by line feeds without duplicates and without any special characters.

Tools > Concatenate Learning Files helps to concatenate existing Learning Files or to create a test file from two Learning files. When the window opens, click **browse** next to **Learning File 1** to select the first Learning file, then click **browse** next to **Learning File 2** to select the second Learning File. In the **Learning File 2 Status** section, select **Learning** if you just want to concatenate selected files or select **Test** if you want to create a Test File. Then, click **OK** and use the save dialog box to select a destination folder and a file name. When you click on **save**, the new file is created.

NOTE1: Variables of the two learning files must be strictly identical otherwise a warning message will be displayed.

NOTE2: Concatenation must be used with care if some objects are present in both original learning files, the concatenated one will contain duplicates. Please refer to the section "Data analysis" for more information about the use of concatenated Learning files.

Tools > Convert to PID Files is not implemented yet.

Help > Online User Guide leads to this web page.

Help > About Plankton Identifier shows *Plankton Identifier* author information.

LEARNING

This step generates a file (Learning file) containing needed information for automatic recognition. It corresponds to a representative sub sample of objects identified by an expert and used as reference in further analysis.

Select Folder Window

When clicking on the **Learning** box, a folder selection window appears (Fig.3). Select a folder which will contain a "Learning Set" corresponding to identified objects (thumbnails) sorted into subfolders (groups) + PID files containing object data.

- To create a new learning set, you must select an empty folder. If necessary, you can create it by clicking on the **new folder icon** (Fig.3: 1) and entering a new name (Fig.3: 2). Then click on the **OK** button (Fig.3: 3).

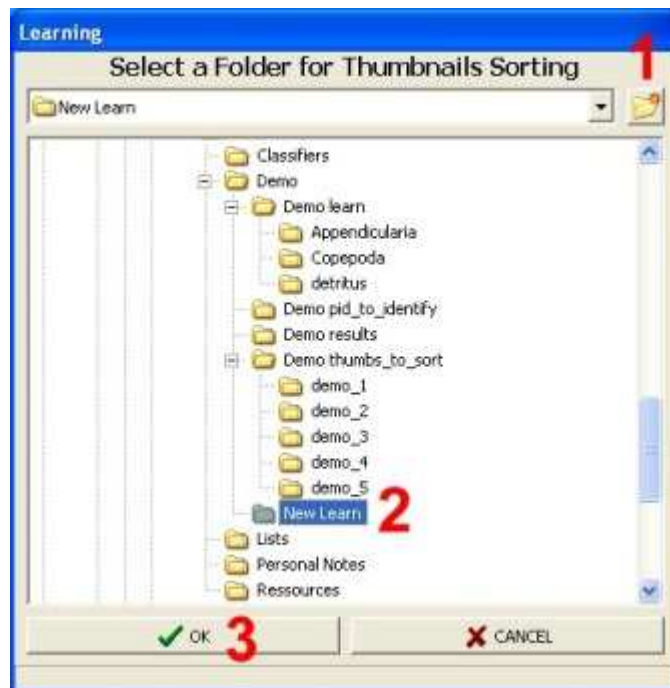


Figure 3: Select Folder Window

- You can also select a folder already containing a learning set. This existing learning set can be an old one you want to modify or a learning set made by other means.

NOTE: if the content of the selected folder does not fit the expected structure or contains invalid data you will not be able to open it and a warning message explaining the problem will appear in red at the bottom of the window.

Learning Window

Once you have selected a valid folder for thumbnails sorting and have clicked OK, a new window appears (Fig.4). The left panel of this window (“Unsorted thumbs”) allows to browse the hard drives to select unsorted objects (thumbnails) and the right panel (“Sorted Thumbs”, empty for a new learning) corresponds to identified objects (thumbnails) sorted into subfolders (groups).

NOTE: Relative size of these different panels can be changed by dragging borders.

Unsorted Thumbnails

In "Unsorted Thumbs" panel, browse the hard drives to open a folder containing thumbnails and their associated PID files. Only thumbnails with a valid name (<Sample Name>_<Item Number>.jpg) will be displayed. If the name is valid but the required data cannot be retrieved, a red question mark will be drawn on the thumbnail (Fig.4: 1) and you will not be allowed to use it. If you place the mouse cursor above the question mark, the reason why required data cannot be retrieved will be displayed (PID file missing, Item missing in the PID file etc.). If the thumbnails appear too small or if you cannot read their names entirely, you can use the “thumbnails size” bar on the left to enlarge them (Fig.4: 2).

NOTE: In \Program Files\Plankton Identifier\Demo\Demo thumbnails_to_sort\, some folders, containing thumbnail images and the corresponding PID file, are available for exercising.

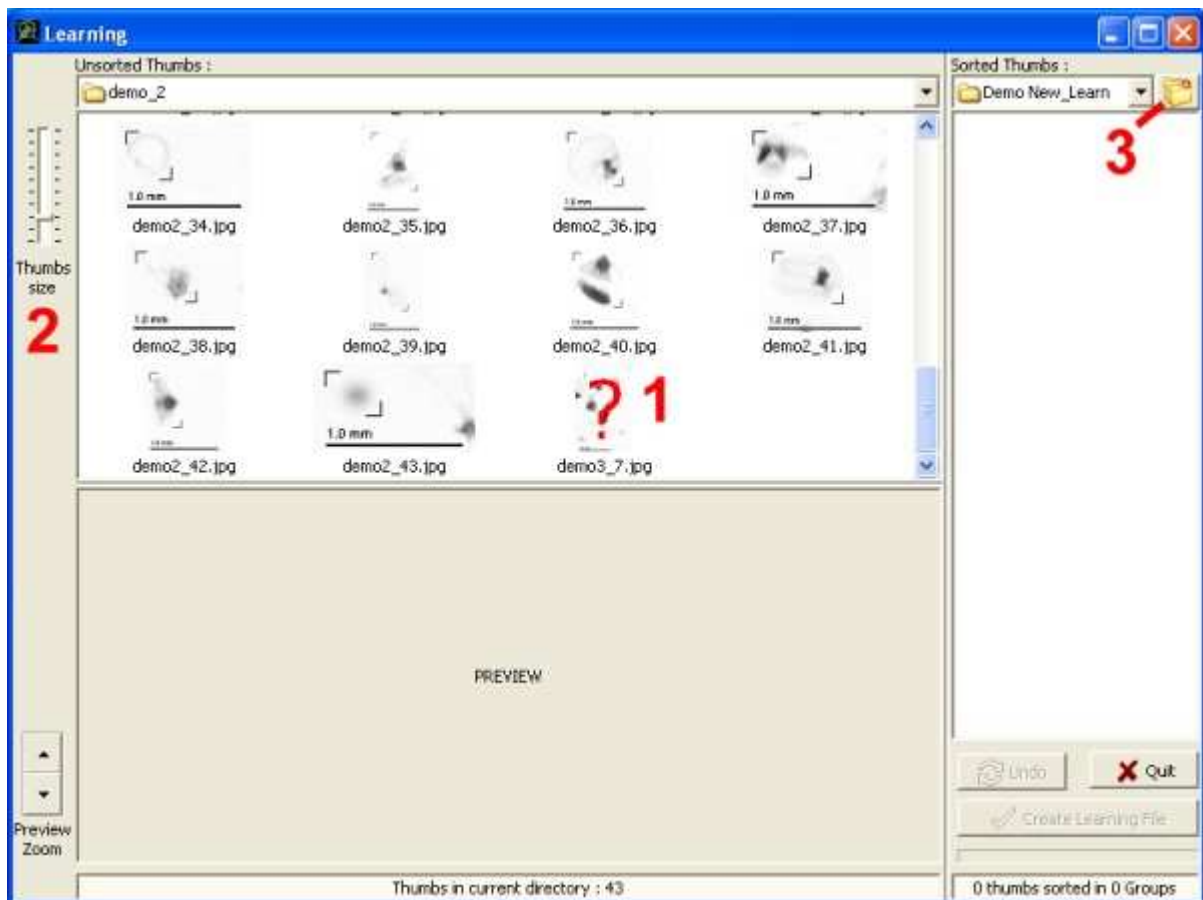


Figure 4: Learning Window

Groups (subfolders) creation

In "Sorted Thumbs" panel, folders must be created to sort the thumbnails according to their identification. Click on the **create new folders** icon (Fig4: 3). It will open a new window which allows the selection of predefined names (Fig.5). If the desired name is not available, you can either select another predefined list or select "new" as name and edit it later in the "Sorted thumbs" panel. When you click **OK**, all the selected names are created as subfolder in the "Sorted thumbs" panel. To edit a folder name, select the folder by a single click, then click again on the name to activate edition.



Figure 5: Create Groups Window

NOTE1: The same folder name cannot be used twice in a learning set. Thus, already used name are not available in the "Create Group Folders" window and any edition of folder name leading to an already existing name will be automatically cancelled.

*NOTE 2: In "Sorted Thumbs" You can open a subfolder and create sub-subfolders in the same way. However, creation of subfolders into subfolder is **not recommended** since further statistical analyses do not use tree structure as valuable information. Only the name of the last folder containing thumbnails will be used as identifier and complex tree structure could create more confusion than advantages.*

*NOTE3: Using the Windows notepad, you can also create your own customized lists or edit an existing list in the "Lists" subfolder of the Plankton Identifier folder. Customized lists will be then available in the Predefined Lists combo-box. A customized list must contain group names separated by line feeds, without duplicates and without any special characters. You can use **Tools>Import Name List** in the main window menu to import your own list.*

Thumbnails sorting

When selecting a thumbnail, a preview appears (Fig.6: 1). You can enlarge that preview with the **Preview zoom** buttons on the left (Fig.6: 2).

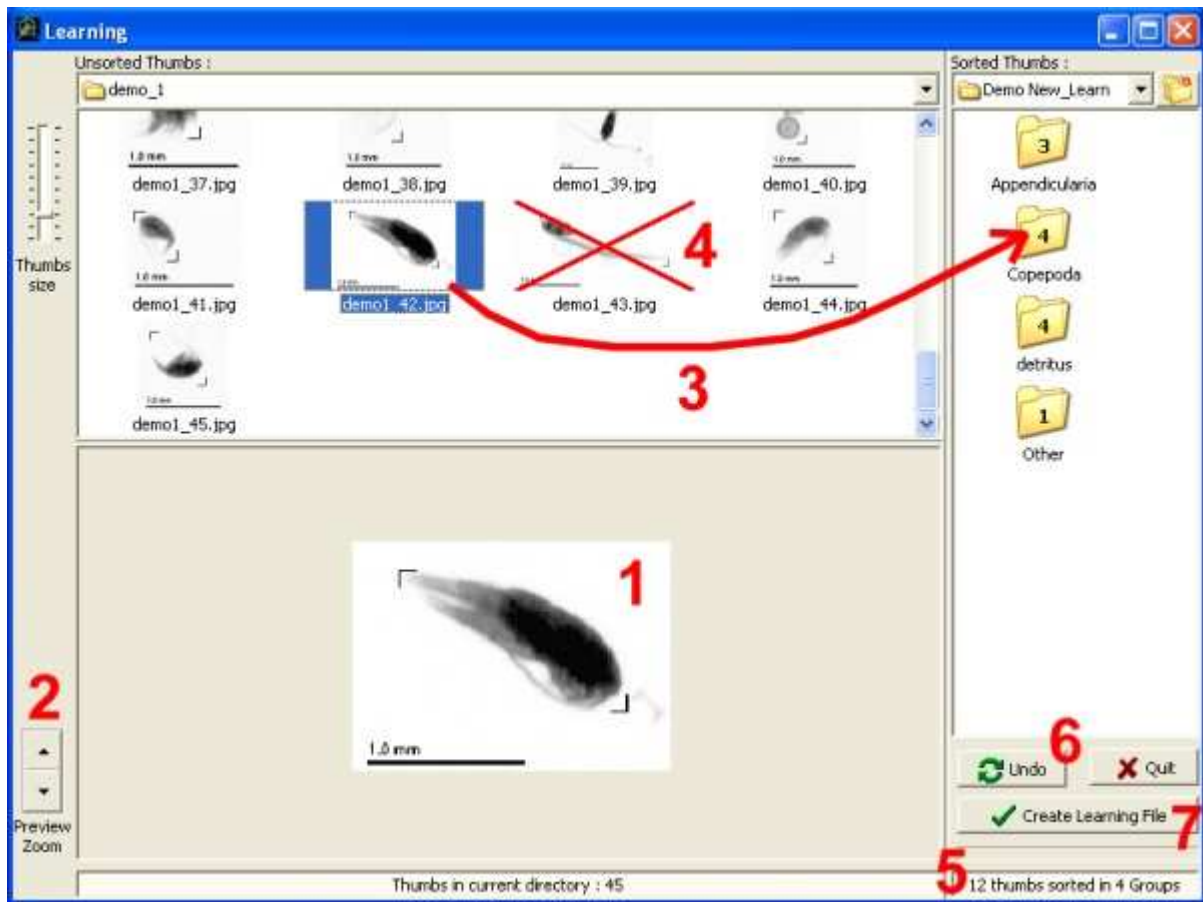


Figure 6: Thumbnails sorting

Drag thumbnails to the corresponding subfolders (Fig.6: 3). When you release a thumbnail on a subfolder, this thumbnail is physically copied (not moved) in the subfolder. The corresponding PID file is also copied in the learning folder but remains invisible to avoid confusion. Once a thumbnail has been used in the current learning set, a red cross is drawn on it (Fig.6: 4) and you will not be allowed to use it anymore.

Selection of several thumbnails at once is possible if you maintain the Ctrl key down during selection. If you want to see the preview of each thumbnail during a multiple selection, then start selection from the down right to the upper left thumbnail position in the panel.

The number of thumbnails in each subfolder is drawn on it and is updated after each drop. The total number of sorted thumbnails, and of those to be sorted, are indicated at the bottom of each panel (Fig.6: 5).

Modify the Learning set

To remove a thumbnail from the Learning set, select it, then use the **DEL key** or right-click on the thumbnail and select "**Delete**" in the popup menu. To remove several thumbnails, select them using the Ctrl or the shift Keys as in Windows shell then proceed as above.

The **DEL key** and the "**Delete**" function of the popup menu can also be used to delete one or several subfolders once they have been cleared out of thumbnails.

To move a thumbnail from one subfolder to another, right-click on the thumbnail and select "**Move to ...**" in the popup menu. To move several thumbnails, first select them using the Ctrl or the Shift Keys.

The popup menu also includes "**Select all**" and "**Invert selection**" functions.

Cancel action

The Undo button (Fig.6: 6) can be used to cancel thumbnails drops, thumbnails deletions, subfolders creations and subfolders deletions as well as subfolder renaming. An unlimited number of actions can be cancelled until you create the Learning File or you quit the Learning Window.

Create a Learning file

Once you consider you have sorted enough objects in each category, click on the **Create Learning File** button (Fig.6: 7). A save dialog box will appear. Indicate a destination folder and a name for the learning file if you do not want to keep the default name (which is Learn_<number>). Click on the **Save** button and the job is done. Then, a dialog box asks you if you want to continue sorting. If you choose "No", the learning window is closed and the main window is enabled.

DATA ANALYSIS

This step generates text files containing the results of the automatic recognition as well as an html report containing data analysis information. Click the Data Analysis button in the main window, a new window opens with different panel sections (Fig. 7):

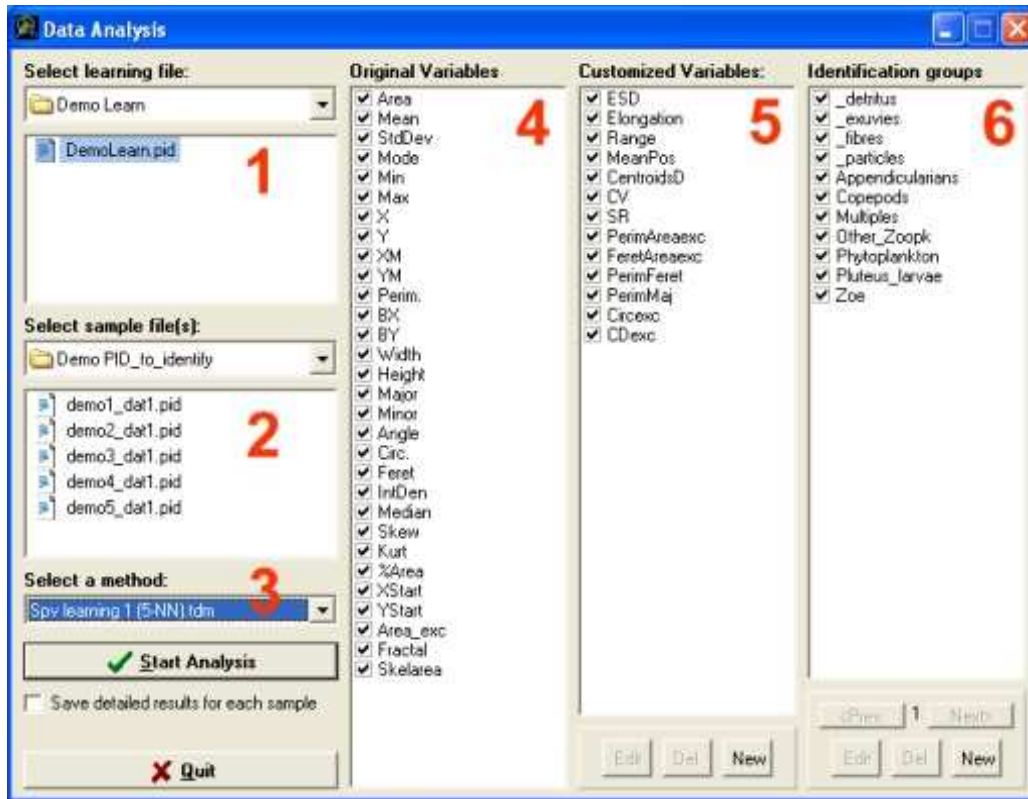


Fig. 7: The Data Analysis window

Select learning file (Fig. 7: 1):

This section allows to browse the hard drives to select the learning file you want to use for analyzing your data.

NOTE: A learning file must be selected to activate other sections.

NOTE 2: A double click on any PID file (Learning or sample) will automatically open it in PID viewer (if installed) or in your text editor (i.e. Windows Notepad). Thus you can easily verify its content if necessary.

Select sample file(s) (Fig. 7: 2):

This section allows to browse the hard drives to select one or more samples (PID files) for which you want to do the automatic object identification. If nothing is selected, analysis will be done with the Learning File only. In that case, a method designed to test the Learning file must be selected.

NOTE 1: To analyse several samples (PID files) at one time, put them in the same subfolder and maintain the 'ctrl' Key down during selection as in Windows shell.

NOTE 2: Variables of selected samples (PID files) must match Learning file variables otherwise an error message will be displayed and analysis is not allowed.

Select a method (Fig. 7: 3):

This section allows the selection of analysis method. Seven supervised learning methods are provided within *Plankton Identifier*:

Spv learning 1 (5-NN): k-nearest neighbour used HVDM distance metric

Spv learning 2 (S-SVC linear): C-SVC from LIBSVM library

Spv learning 3 (S-SVC RBF): C-SVC from LIBSVM library

Spv learning 4 (Random Forest): (Breiman, 2001)

Spv learning 5 (C4.5): Decision tree algorithm (Quinlan, 1993)

Spv learning 6 (Multinomial Logistic Regression): Multinomial Logistic Regression with a Ridge estimator (S.le Cessie & J.C.van Houwelingen, 1992)

Spv learning 7 (Multilayer Perceptron): Multilayer Perceptron neural network

In addition, the following selections are possible:

Cross-validation 1 to 7 : Evaluate the accuracy of one of the seven algorithms using a re-sampling technique. The original learning set is randomly partitioned into two subsets of the same size. One of the subsets is used as learning set to build the predicting model, and the remaining subset is retained as the validation data for testing the model. Each subset will be used once as learning set and once as testing set. The two results from the testing sets are then averaged to produce a single estimation of the predicting model performance. The cross-validation process is repeated 5 times and the average error rate of the 5 cross-validations is computed in a confusion matrix.

To use Cross-validation :

1. Select a learning file (check variables and identification groups)
2. Do NOT select any sample file
3. Select the Cross-validation method corresponding to your preferred algorithms
4. Press Start Analysis.

Test 1 (7 methods): Compares the seven supervised learning algorithm accuracy performances on a predefined test file. Before using Test 1, a special file (test file) must be created with the tool **Concatenate Learning Files** (main window menu). Rather than to use two distinct files for the learning and the testing set, we prefer join them together in a single file and use the column Status to indicate the role that each observation must play. The learning data (Status = Learning) will be used in the learning process to build the seven classification models (or classifiers) whereas the testing data (Status = Test) will be used to obtain an unbiased error rate evaluation.

To use Test1:

1. Select the file you have just created as data source
2. Do NOT select any sample file
3. Select the Test 1 method
4. Press Start Analysis.

NOTE 1: The learning and testing set are assumed to be representative of the same set of observations.

NOTE 2: Usually , the larger is the learning set, the better is the classifier and the larger is the testing set the more accurate the predictive accuracy, or error estimation.

Export to text file (no analysis): generates a text file containing a concatenation of selected learning file and sample file(s), with all the original variables + selected customized variables + edited groups but without any recognition. This text file can be then imported in any data mining software for analysis.

Original variables (Fig. 7: 4):

This section shows the variables available in the selected Learning File. You can enable or disable the variables to be used for the analysis. Disabled original variables will be ignored but will not be removed from result files.

Customized variables (Fig. 7: 5):

This section helps to create new variables from existing original ones. Thirteen customized variables that you can enable or disable are already available when you install *Plankton identifier*. Disabled customized variables will be ignored and will not appear in result files. If an existing customized variable cannot be calculated from original variables which are currently available, this variable is automatically disabled and appears in grey.

- To edit an existing customized variable, select it and click on **Edit** to open the Customized Variable window (Fig. 8).
- To erase definitely an existing customized variable, select it and click on **Del**.
- To create a new customized variable, click on **New** to open the Customized Variable window
 1. Give a Name to the new variable (it must be different from existing variables)
 2. Enter a formula in the field "Operation"
 3. Press OK.

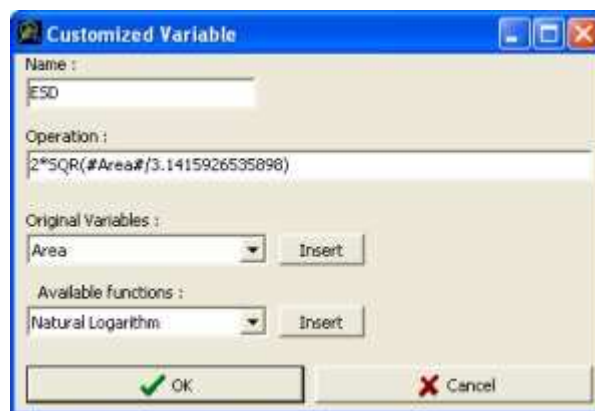


Figure 8: Customized Variable window

*NOTE: To write a formula, keyboard operator (e.g. +, -, /, *, ^), standard brackets and numbers can be used as usual. For constant (e.g. numbers), the decimal separator depends on your regional setting. To insert one of the original variables, select it using the "original variable" box, and press **Insert**. Variable name will appear surrounded by "#" (do not remove these marks without removing the whole variable name). To insert a function, select it using the "Available functions" box, and press **Insert**. We recommend to look at already existing customized variables (use **Edit**) to see how to built new formula.*

Identification Groups (Fig. 7: 6):

This section shows the Identification groups defined in the selected learning file. You can enable or disable groups to be used or not for the analysis. Disabled groups will be ignored and will be removed from result files. This section also allows to create new groups by grouping existing groups. The original group list (#1) cannot be deleted or edited.

- To create new groups, click on **New** to open the Groups Edition window (Fig. 9).



Fig. 9: The Groups Edition Window

- To change original group names (Fig. 9: 1):
 1. Select the name in the Modified Group Names section (Fig. 9: 2).
 2. Write a new name for this group or click on the button (Fig. 9: 3) and select a name in the list. Groups with the same new name will be associated.
 3. When you finish to modify the group names, press **Done**

NOTE: You can create as many lists of Identification groups as you want. The list used in the analysis will be the current list (i.e., the one visible in the Data Analysis window when you will click on Start Analysis). You can use the <Prev. and Next> buttons to select the list you want to use for the analysis. In the result files, the original names will appear in the column "Ident" and the new names in the column "Ident2", if used.

NOTE 2: New groups definition will apply even if you select another Learning file having different original group names. In this case, only group names already encountered will be redefined.

- To edit an Identification Groups list, use the <**Prev.** and **Next**> buttons to select the list then click **Edit** to open the Group Edition window and proceed as above.
- To delete the current Identification Groups list, press **Del**.

Launch analysis

When all files, variables and identification groups are selected, click on the **Start Analysis** button. A save dialog box will appear. Indicate a destination folder and a name for the result file if you do not want to keep the default name (which is Analysis_ <number>). Click on the **Save** icon and the analysis is launched.

Once the analysis is finished (it can take several minutes depending on the sample size and the selected method), the results and the html report are saved in the selected destination folder. A dialog box asks you if you want to quit the Data analysis window. If you choose “Yes”, the Data Analysis window is closed and the main window is enabled.

Compared to the original PID file, the result file(s) contain(s) several new columns:

1. Columns corresponding to customized variables if used
2. One column (Ident) containing group names in the Learning file
3. One column (Ident2) containing modified group names if used
4. One column (Status) containing object status (possible values are Learning, Sample and Test)
5. One column (**Pred_xxxx**) containing predicted identification.

IMPORTANT: Columns Ident and Ident2 of the result(s) file(s) correspond to visually identified object when the object status is "Learning" but are meaningless when the object status is "Sample" (just filled with the first group name of the Learning File). Look at the last column (Pred_xxxx) for predicted object identification.

*NOTE: If you want to create a separate result file, keeping the original Header, for each sample (PID file) used in the analysis, enable the checkbox **Save detailed results for each sample** under the Start Analysis button before launching the analysis. These files are created in addition to the main result file with names as follows: <Result File Name>_<sample name>.txt*

SHOW REPORT

This step facilitates the access to the result files and to the html reports of your analysis. Then, it is simple to export results towards an edition software (like Excel©) for subsequent processing, or to print the results.

Click on the **Show Report** button. In the Show Report window (Fig. 10), the last folder successfully used as destination folder for results is automatically opened and the last result file appears in bold characters (Fig.10: 1). It is possible to browse the hard drive to retrieve results of previous analysis.



Figure 10: Show Report window

- Double click on the result file to open it in the default text editor (i.e. notepad)
- Select the result file and click **OK** (Fig.10: 2) to open the corresponding html report in the default html browser (i.e. internet explorer)

NOTE: If detailed results for each sample have been saved separately, all the files can be opened in the default text editor but only the main result file provide direct access to the html report.

The Html report (Fig. 11) is the one of Tanagra. The left panel shows the data analysis diagram. When clicking on one item of this diagram, corresponding information are shown in the right panel. In Figure 11, for example, a click on "Univariate discrete Stat 1" on the left shows the distribution histogram of predicted object identifications on the right. Copy-Paste can be used to export this information towards another software. Please refer to Tanagra documentation for more information about diagram items and results interpretation.

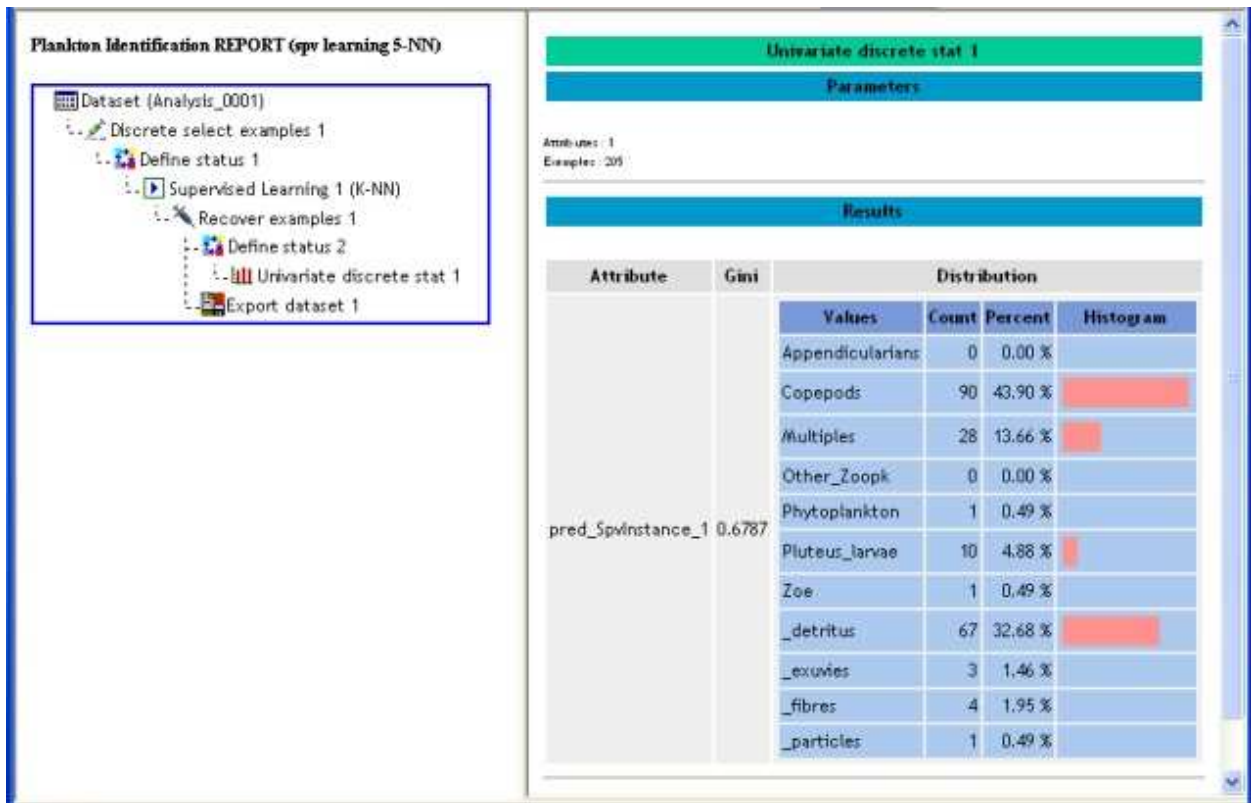


Figure 11: Html report

APPENDIX

File formats and File names

To work with the current version of *Plankton Identifier*, PID files and thumbnails are needed. Both are generated from the original image by other software such as ZooProcess.

PID Files

A PID file collects object information about one image corresponding usually to one sample in the real world. Each PID file is basically a text file very similar to an initialization (.ini) file. The first line contains a signature (PID) and is followed by different sections. Each section starts with a section declaration, which corresponds to section name surrounded by square brackets [].

The first sections contain information about the sample, sample processing and image acquisition etc. In these sections, forming the header, each information item is made up of an item name, equals sign ('='), and a value.

The last section, [Data], contains object measurements. **Only this last section is used by the current version of *Plankton Identifier*.** The first line of this section contains column names (i.e. variable names) separated by semicolons (;). The other lines contain corresponding data (one line per object) also separated by semicolons.

The first column must correspond to a unique number identifying the object in the sample, and the second column must correspond to the sample name. Other columns correspond to object measurements. The number of values must be the same in all lines.

NOTE: variables locating objects in the original image are required for post processing including visual validation, but not for automatic recognition.

More information about PID files is available in the ZooProcess manual.

Thumbnails

Each thumbnail corresponds to the image of one object in the original image. The jpeg format has been chosen because thumbnails are used for visualization only and not for measurements. In such conditions, image degradation due to jpeg compression is not a problem and jpeg format preserves a lot of disk space.

To be used with *Plankton identifier*, the name of each thumbnail must contain a prefix, identifying the corresponding PID file, and a suffix, identifying the object in the [Data] section of the PID file, separated by an underscore “_”. If several underscores are present in thumbnail name, only the last one is considered as separator.

The prefix must correspond exactly to the **beginning of the name** of the corresponding PID file but the end of PID file name can include some additional comments. For example, a thumbnail named **DEMO1_24.jpg** can correspond to a PID file named **DEMO1(zoopk).pid**. Of course, equivocal situations must be avoided:

Example 1:

Bad situation: **DEMO1_24.jpg** coexists with **DEMO1_Test1.pid** and **DEMO1_test2.pid**.

Solution: Use **DEMO1_Test1_24.jpg** or **DEMO1_Test2_24.jpg** as thumbnail name.

Example 2:

Bad Situation: **DEMO1_24.jpg** coexist with **DEMO1(zoopk).pid** and **DEMO10(phytopk).pid**.

Solution: Change **DEMO1(zoopk).pid** to **DEMO01(zoopk).pid** and use **DEMO01_24.jpg** as thumbnail name.

The suffix corresponds to the object number in the first column of the [data] section of the PID file, meaning that the same number must not be use twice in the same PID file.

Before sorting (see LEARNING), thumbnails are expected with their associated PID file in the same folder by *Plankton identifier*.

Learning Files

Learning Files are generated and used by *Plankton identifier*. These are text files with a .pid extension but differ from original PID files (see above). The first line contains another signature (LEARNING) to avoid confusion. The second line contains column names (i.e. variable names) separated by semicolons (;) and the other lines contain corresponding data (one line per object) also separated by semicolons. Compare to the [Data] section of original PID files, a Learning File contains data of identified objects only as well as two additional columns. The first additional column is named "Ident", and contains the name attributed to the object by an expert. The second additional column is named "Status", and contains the word "Learning" in order to differentiate the Learning part in further concatenated files.

Test files are special Learning File resulting from the concatenation of two Learning Files, one receiving the word "Test" rather than "Learning" in the column "Status". Test file associated to specific tdm are useful to evaluate the efficiency of a Learning File.

Folders organization

In order to preserve flexibility and control by the users, many files and folders are created during the different process and can be accessed using Windows shell. If you do not take care of files and folders organization, the situation will quickly become confusing.

Recommendations

The use of **one folder per project** is highly recommended with **at least** the following subfolders:

1. One subfolder per set of thumbnails extracted from one original image with the corresponding PID file
2. One subfolder per Learning Set (compulsory, see below)
3. One or more subfolder(s) for PID files (samples) to be analyzed
4. One or more subfolder(s) for data analysis results.

See the "Demo" folder provided with *Plankton Identifier* for an example.

Learning Set

A “Learning Set” corresponds to one folder containing PID files and thumbnails sorted into subfolders. In opposition to folder containing unsorted thumbnails, in a “Learning Set” PID files are not stored in the same subfolder as thumbnails but in the root folder. For this reason, if you try to use a “Learning Set” as source of unsorted thumbnails, these later will appear with red question marks in the *Plankton Identifier* “Learning” window (and even with question marks and crosses if copies of these thumbnails are already used in the current Learning Set). Thus, unintended cross-sorting is impossible.

The name of the subfolder containing a thumbnail is used to fill the column “Ident” of the corresponding object when creating a Learning File. This is the reason why all the different subfolders must have different names even as sub-subfolder.

A “Learning Set” is valid only if all the thumbnails it contains have the corresponding PID file in the root directory (See the “File formats and File Names” section for information about naming convention). A “Learning Set” can contain additional files such as text files with some comments, Learning Files or unused PID files but jpeg images other than thumbnails are not allowed.

Build customized analysis method (tdm)

Plankton identifier uses Tanagra tdm files to register data analysis methods. These tdm files are stored in the “Classifiers” subfolder of the *Plankton Identifier* install directory. To add a customized data analysis method to *Plankton Identifier* , a tdm file must be created using Tanagra then imported in this subfolder. Because it is not *Plankton identifier* purpose to reproduce all Tanagra functionality, a tdm file compatible with *Plankton Identifier* has some limitations and must comply with the following rules:

- It must contain at least a component “*Define status 1*” and a component “*Export data set 1*”.
- “*Define Status 1*” must contain measurements as input (continuous variables) and “*Ident*” as target (discrete variable).
- If others “*Define status*” are used, they cannot make reference to measurements as input or as target (but they can make reference to discrete variables, or to outputs of previous components)
- Other components such as “*Continuous select examples*” cannot make reference to measurements or to “*Ident*” (Only “*Define status*” components can make reference to “*Ident*”)

Since a tdm file is basically a text file, it can be created or edited using a simple text editor such as the notepad. However, the simplest way to create a new tdm file for *Plankton Identifier* is the following:

- In the Data Analysis window of *Plankton Identifier*, select a learning file and a sample file. Select “Export to text file (no analysis)” in the select method box, verify that the active Identification groups list is the first one, then save the file by clicking on start analysis.
- Launch Tanagra and select File>New...
- In the Dataset field, select the previously created text file then click OK.
- Construct your data analysis diagram (see Tanagra documentation) having in mind the rules described above.
- Select File>Save as... and save the diagram as “Text data mining diagram” in a temporary folder with an easy to understand name (this name will appear in the select method box of *Plankton Identifier*)
- Go back to *Plankton Identifier* and select “Import Tanagra Data Mining diagram” in the “Tools” menu of the main window.
- Select the newly created tdm file and click Open... The tdm file is copied in the appropriate subfolder and is now available in the select method box of the Data Analysis window.

F.A.Q.

How many groups can I create ?

You can create as many groups as you want. Actually, you can even take advantage to create very specific groups (i.e., to species level). If the classifier model is unable to distinguish all the groups you have defined, you will be able to test different group associations to improve classifier performances using the appropriate tools of the Data Analysis section.

How many objects must be sorted ?

The larger is the learning set, the better is the recognition accuracy. However, from our experience (zooplankton recognition using Zooscan and Zooprocess), 100 objects per group is a minimum to get a good classifier model.

Why data analysis always fails with my computer?

For Plankton Identifier version below 1.2.6, It is because the decimal separator of your Windows configuration is different from the decimal separator used in PID files. Go to start menu -> parameters -> Control panel -> Regional settings and change decimal separator to '.' or update to version 1.2.6.

Why Random forest method is not working with my computer whereas other methods work ?

It is because your Tanagra version is too old. Update to Tanagra 1.4.12 or above.

Why Plankton Identifier sometimes crashes when I try to analyse my own PID files ?

It is probably because at least one of your PID files is partially corrupted (i.e. do not strictly fit the expected format). To solve this problem, open your PID files with PID viewer and verify their content.

What are the different variables used by Plankton Identifier ?

Original variables are those measured by the image analysis software that generates PID files. Although many image analysis softwares could be used with Plankton Identifier, ImageJ is currently the most commonly used one. This is the reason why customized variables provided with Plankton Identifier are based upon original variables measured by ImageJ and Zooprocess. The list below describes briefly these different variables:

NOTE: Those in italic being meaningless for distinguishing between groups, they MUST be deselected before starting an analysis

Original variables from ImageJ:

Area: Surface of the object in square pixel.

Mean: Average grey value within the object; this is the sum of the grey values of all the pixels in the object divided by the number of pixels

StdDev: Standard deviation of the grey value used to generate the mean grey value

Mode: Modal (most frequently occurring) grey value within the object

Min: Minimum grey value within the object (0 = black)

Max: Maximum grey value within the object (255 = white)

X: X position of the centre of gravity of the object (can be used in customized variables, do not use it directly as a measurement)

Y: Y position of the centre of gravity of the object (can be used in customized variables, do not use it directly as a measurement)

XM: X position of the centre of gravity of the grey level in the object (can be used in customized variables, do not use it directly as a measurement)

YM: Y position of the centre of gravity of the grey level in the object (can be used in customized variables, do not use it directly as a measurement)

Perim: The length of the outside boundary of the object

BX: X coordinate of the top left point of the smallest rectangle enclosing the object (used to extract thumbnails, not really a measurement)

BY: Y coordinate of the top left point of the smallest rectangle enclosing the object (used to extract thumbnails, not really a measurement)

Width: Width of the smallest rectangle enclosing the object (used to extract thumbnails, not really a measurement)

Height: Height of the smallest rectangle enclosing the object (used to extract thumbnails, not really a measurement)

Major: Primary axis of the best fitting ellipse to the object.

Minor: Secondary axis of the best fitting ellipse to the object.

Angle: Angle between the primary axis and a line parallel to the x-axis of the image (used to get object positioning, not really a measurement)

Circ: Circularity = $(4 * \text{Pi} * \text{Area}) / \text{Perim}^2$; a value of 1 indicates a perfect circle, a value approaching 0 indicates an increasingly elongated polygon. (it is the reverse of compactness)

Feret: The maximum Feret's diameter, i.e., the longest distance between any two points along the object boundary.

IntDen: Integrated density. This is the sum of the grey values of the pixels in the object (i.e. = Area*Mean)

Median: Median of the grey value used to generate the mean grey value.

Skew: The third order moment about the mean. It is the measure of lack of symmetry. The skewness for a normal distribution is zero. Negative values for the skewness indicate data that are skewed left and positive values indicate data that are skewed right.

Kurt: The fourth order moment about the mean. It is a measure of whether the data are peaked or flat relative to a normal distribution. Positive kurtosis indicates a peaked distribution and negative kurtosis indicates a flat distribution.

%area: Surface of holes in percentage.

XStart: X coordinate of the top left point of the image (used to locate object, not a measurement)

YStart: Y coordinate of the top left point of the image (used to locate object, not a measurement)

Area_exc: Surface of the object excluding holes in square pixel (=Area*(1-(%area/100))

Mean_exc: Average grey value excluding holes within the object (= IntDen /Area_exc)

Fractal: Fractal dimension of object boundary

Skelarea: Surface of skeleton in square pixel.

Customized variables implemented by Plankton Identifier:

ESD = Equivalent Spherical Diameter = $2 * \text{SQR}(\text{Area} / \text{Pi})$

Elongation = Major / Minor ('ellipse' elongation)

Range = Max - Min

MeanPos = (Max - Mean) / Range

CentroidsD = $\text{racine}((\text{XM} - \text{X})^2 + (\text{YM} - \text{Y})^2)$; distance between the centroid and the centroid of mass

CV = $100 * (\text{StdDev} / \text{Mean})$

SR = $100 * (\text{StdDec} / (\text{Max}-\text{Min}))$

PerimAreaexc = Perim / Area_exc

FeretAreaexc = Feret/Area_exc

PerimFeret = Perim / Feret

PerimMaj = Perim / Major

Circexc = $(4 * \text{Pi} * \text{Area_exc}) / \text{Perim}^2$

CDexc = $(\text{CentroidsD})^2 / \text{Area_exc}$

What is the best supervised learning method ?

The best method to use depends on your samples. Use a test file and the method 'Test 1' to compare supervised learning method performances at once, then select the one that gives the best results in your case.

How can I verify that my preferred groups are well distinguished ?

Use cross-validation and look at the confusion matrix. If two groups are not well distinguished but their association make sense, try to put them together in a larger group and run cross validation again.

How to read a confusion matrix ?

A confusion matrix is a matrix showing the actual versus predicted classifications. A confusion matrix is of size $k \times k$, where k is the number of classes. The following confusion matrix is for $k = 2$ classes:

	Predicted Positive	Predicted Negative	
Actual Positive	TP	FN	n+
Actual Negative	FP	TN	n-
	TP + FP	FN + TN	N

Given a classification model (or classifier) and one instance, there are four possible outcomes:

- If the instance is positive and it is classified as positive, it is counted as a *true positive* (TP)
- If the instance is positive and it is classified as negative, it is counted as a *false negative* (FN)
- If the instance is negative and it is classified as negative, it is counted as a *true negative* (TN)
- If the instance is negative and it is classified as positive, it is counted as a *false positive* (FP).

Given a classifier and a set of instances (the learning set or the testing set) a two by two confusion matrix can be constructed. Several common performance metrics can then be calculated from it.

Accuracy rate (1 - Err. rate): $(TP + TN) / N$

The rate of correct predictions made by the model over the data set (N). It corresponds to the numbers along the major diagonal.

*NOTE 1 : The **re-substitution accuracy rate** corresponds to the accuracy made by the model on the learning set . The accuracy on the learning data is NOT a good indicator of performance on future data since it does not measure any not yet seen data. One way to overcome this problem is to estimate accuracy by using an independent testing set that was not used at any time during the learning process. More complex accuracy estimation using re-sampling techniques, such as cross-validation, are commonly used, especially with data sets containing a small number of instances.*

NOTE 2 : The use of accuracy to evaluate a model assumes uniform costs of errors and uniform benefits of correct classifications.

Error rate (= 1 - Acc.rate): $(FP + FN) / N$

The rate of incorrect predictions made by the model over the data set (N). It corresponds to the numbers off the major diagonal (i.e., the confusion).

True positive rate (Recall, Sensitivity): TP / n_+

The rate of positives (TP) correctly classified as positive by the model over the positive instances (n_+) of the data set (N).

False positive rate: FP / n_-

The rate of negatives (FP) incorrectly classified as positive by the model over the negative instances (n_-) of the data set (N).

Specificity (=1- false positive rate): TN / n_-

The rate of negatives (TN) correctly classified by the model over the negative instances (n_-) of the data set (N).

Precision: $TP / (TP + FP)$

The rate of positives in the positive predicted class.