Supervised DNA barcodes species classification: analysis, comparisons and results

Tutorial

Emanuel Weitschek, Giulia Fiscon, and Giovanni Felici

Citations

If you use this procedure please cite:
Weitschek E, Fiscon G, and Felici G
Supervised DNA Barcodes species classification: analysis, comparisons and results
BMC BioData Mining

Contents

Introduction

About the tutorial

This manual is intended for all users who want to learn how to use the Weka Machine Learning software and the special FASTA sequences converter for DNA Barcodes classification. It is process-based and explains how to achieve a specific objective on a step-by-step basis.

Supervised Machine Learning

The goal is to assign an unknown specimen to a known species starting from its DNA Barcode sequence.

The supervised machine learning classification problem may be formulated in the following way. Given a reference library composed of DNA Barcode specimen sequences of known species and a collection of unknown DNA Barcode sequences (query set) recognize the latter into the species that are present in the library

To obtain reliable results:

- the query set has to contain only specimens from the same species that are present in the reference library;
- the reference set has to contain a sufficient number of specimens sequences for each species (our experiments show that at least 4 specimens per species are necessary to obtain a reliable classification rate);
- the sequences of each species in the reference set have to include possibly all the nucleotide polymorphisms (variations).

The user has to provide as input a training set (reference library) containing specimens with a priori known species membership.

Based on this training set, the software computes the classification model.

Subsequently, the classification model can be applied to a test set (query set) which contains specimens that require classification.

The test set can contain query specimens with unknown species membership or, alternatively, specimens that also have a priori known species membership, allowing verification of the specimen classifications.

Supervised Machine Learning procedure steps

The following steps are necessary to perform a DNA Barcodes sequences classification with the suggested supervised machine learning procedure. These steps are going to be thoroughly explained in the next sections of this tutorial.

1. Sequences acquisition;
2. Sequences alignment;
3. Reference and query split;
4. Software download and installation;
5. Sequences conversion;
6. Classification

DNA Barcode sequences download

To run the experiments, firstly you have to acquire the DNA Barcode FASTA sequences, for example you can download them from dmb.iasi.cnr.it/supbarcodes.php and BOLD system (www.boldsystems.org/).

Sequences alignment

If not already performed by the data download system, please align the acquired sequences obtaining DNA Barcode sequence alignments in the standard FASTA format (query and reference).

Reference and query split

The sequences have to be split in reference and query sets. See the supervised machine learning section above for further explanation.

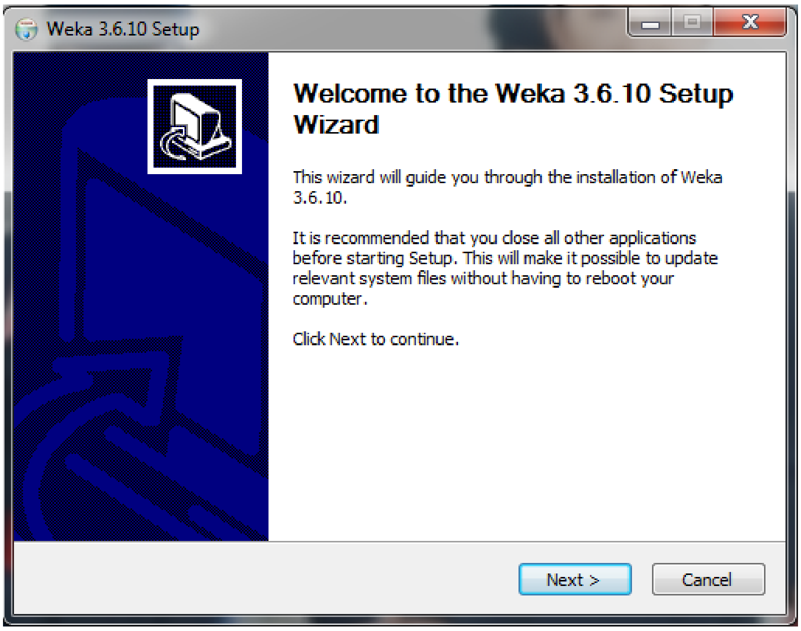
Installation

JAVA

Weka and the special FASTA converter strictly require a working JAVA Virtual Machine (VM) installed. Before going on download and install the free Oracle JAVA Runtime Environment from www.java.com/getjava/.

Weka

You can download and install Weka from www.cs.waikato.ac.nz/ml/weka/downloading.html.
Several versions for the most common operating systems are available, please chose Windows x86 if you have Windows 32 bit, Windows x64 for Windows 64 bit, MacOsX, or Linux. Please chose the versions without Java VM according to your operating system.
Then follow the install wizard by agreeing to the license agreement and by clicking "next" to the following steps.



Weka 3.6.10 Setup

Welcome to the Weka 3.6.10 Setup Wizard

This wizard will guide you through the installation of Weka 3.6.10.

It is recommended that you close all other applications before starting Setup. This will make it possible to update relevant system files without having to reboot your computer.

Click Next to continue.

Weka installation wizard

Fasta to Weka Converter

You can download and unzip an integrated multi-platform (Windows, Linux and MacOS) Java software FASTA to Weka converter from dmb.iasi.cnr.it/supbarcodes.php ("fasta2weka.zip"), that transforms the DNA Barcode FASTA sequences to the suitable ARFF Weka format.

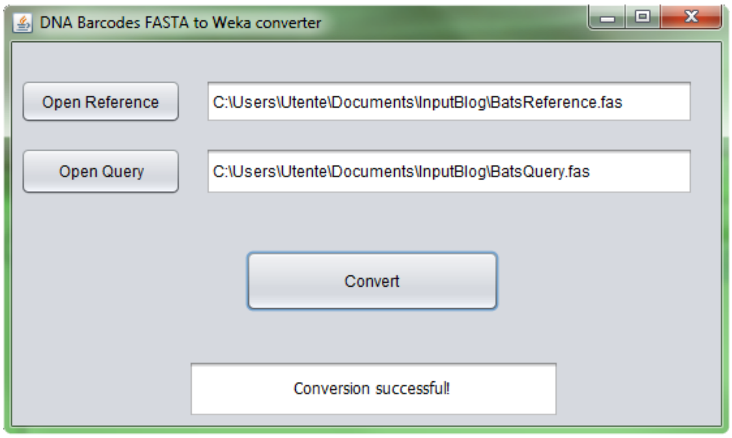
Executing the programs

1. Fasta2Weka Converter

Go to the directory where you extracted the Fasta2Weka archive and execute fasta2weka.bat for Windows and fasta2weka.sh for Linux/MacOs.

Start screen

Once you executed the converter, the main screen of the converter will appear.



C:\Users\Utente\Documents\InputBlog\BatsReference.fas

C:\Users\Utente\Documents\InputBlog\BatsQuery.fas

Conversion successful!

Converter start screen

Input

The input of Fasta2Weka converter is the standard FASTA format of the DNA Barcode Sequences. The heading line of each sequence is composed by the starting character ">" and the "specimen id" and "species name field" separated by the pipe character "|" (eg: ">E3434243 | squalus edmundsi").

Run Fasta to Weka Converter

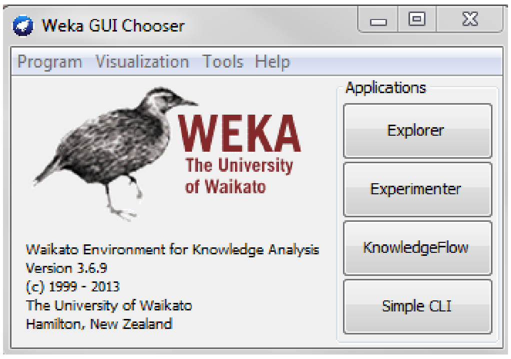
To succeed the conversion from FASTA to ARFF format you have to:

- click the **Open Reference** button and select the reference sequence (FASTA) to be converted;
- click the **Open Query** button and select the query sequence (FASTA) to be converted (optional if you want to convert only the reference set) ;
- click the button **Convert** to start the conversion of both reference (and query) sequences.

If the conversion succeeds , the message "Conversion Successful" will be displayed.
The output files will be saved in the same directory of the input files.

Weka

Go to the start menu and execute the Weka GUI. The following screen will appear. Please choose the “Explorer” Application.

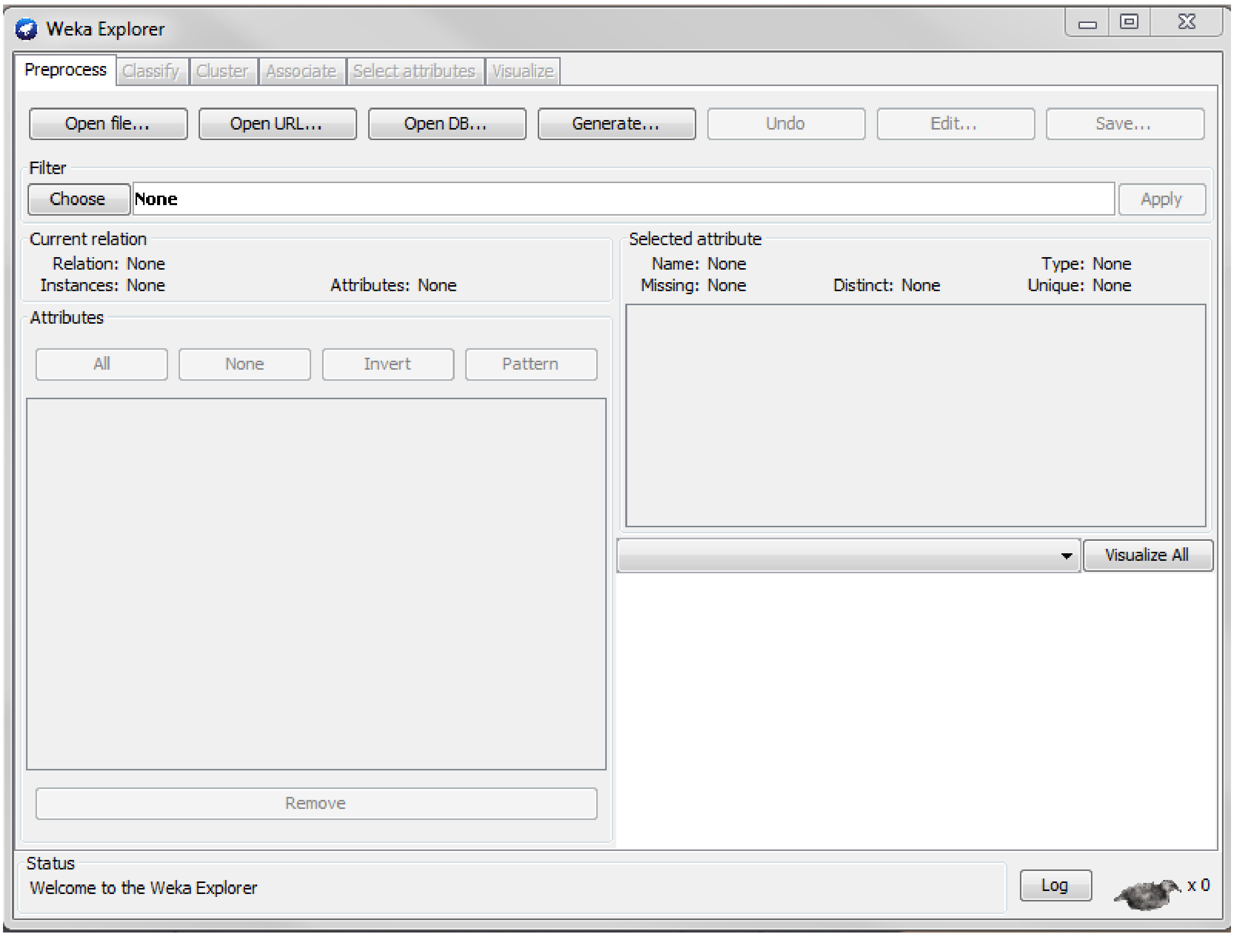


Waikato Environment for Knowledge Analysis
Version 3.6.9
(c) 1999 - 2013
The University of Waikato
Hamilton, New Zealand

Weka GUI start window

Weka Explorer Start screen

The main screen in Weka Explorer is the starting point for all actions. The **Weka Explorer start screen** has different sections: Preprocess, Classify, Cluster, Associate, Select Attributes, and Visualize. Please go to the “Preprocess” section in order to load the datasets to be analyzed.



Filter: None

Relation: None
Instances: None
Attributes: None

Name: None
Type: None
Missing: None
Distinct: None
Unique: None

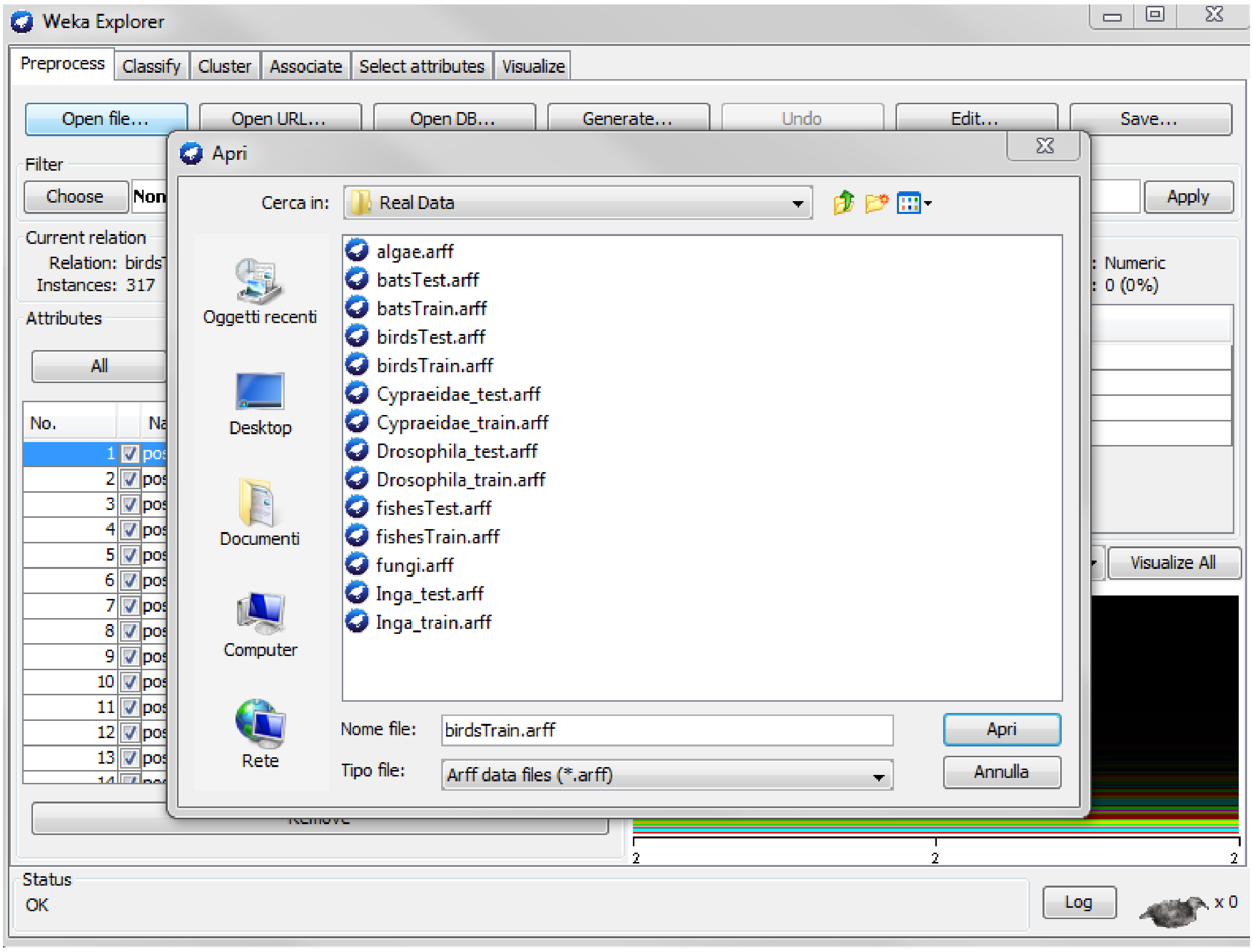
Welcome to the Weka Explorer

Weka Explorer start screen

Input

The input format of Weka are the DNA Barcode sequences in ARFF format. We remind that the common FASTA DNA Barcodes sequences need to be converted. The “Open File” button allows user to select the ARFF datasets to be opened.

- Click the **Open File** button in the Weka Explorer start screen and the **file selection window** appears.



Cerca in: Real Data

Nome file: birdsTrain.arff

Tipo file: Arff data files (*.arff)

Relation: birdsT

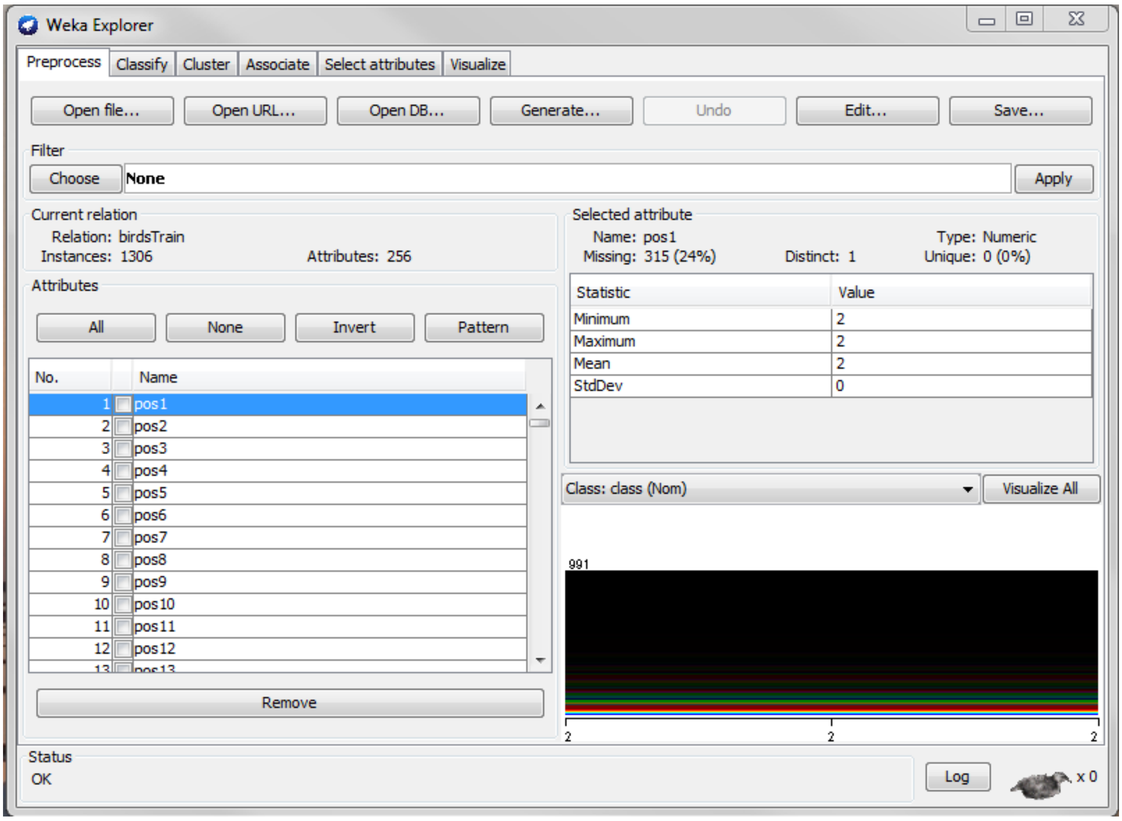
Instances: 317

Status

OK

File selection window

- Select the .arff file of reference set (training set) that you want to open and click the Open button: the arff file of the reference set is now open and the number of instances and attributes is displayed.



Filter: None

Current relation
Relation: birdsTrain
Instances: 1306
Attributes: 256

Selected attribute
Name: pos1
Missing: 315 (24%)
Distinct: 1
Type: Numeric
Unique: 0 (0%)

Statistic	Value
Minimum	2
Maximum	2
Mean	2
StdDev	0

Attributes

No.	Name
1	pos1
2	pos2
3	pos3
4	pos4
5	pos5
6	pos6
7	pos7
8	pos8
9	pos9
10	pos10
11	pos11
12	pos12
13	pos13

Class: class (Nom)

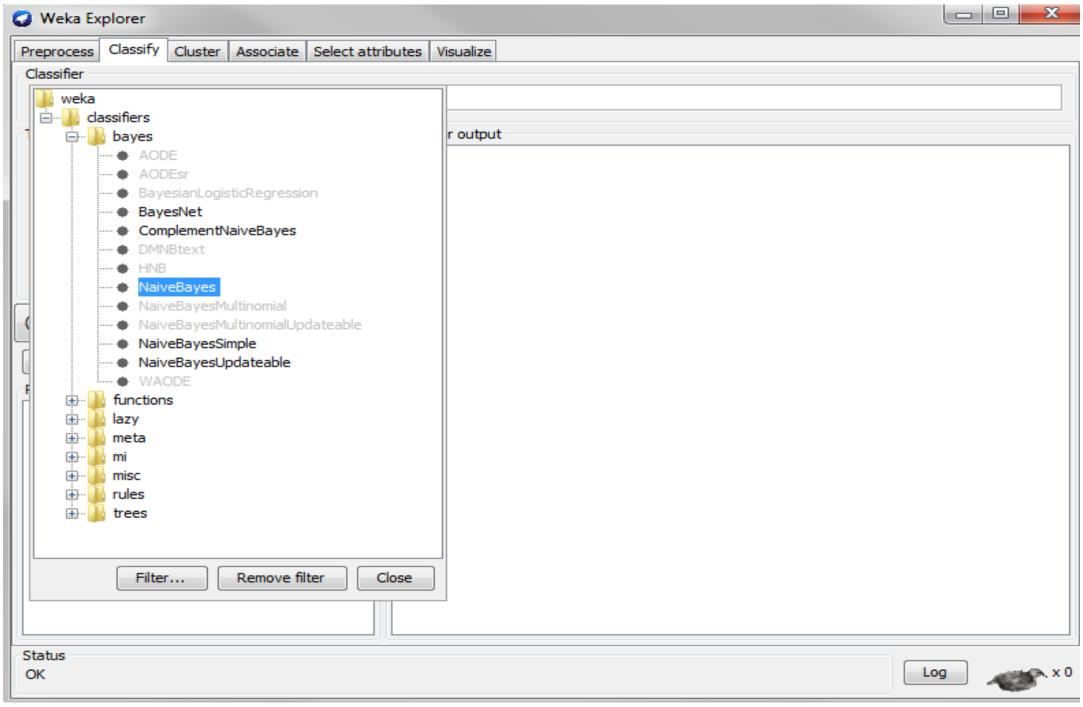
Status
OK

Input selected file window

Classifiers

Once you open a reference file, the section “Classify” is activated and the “choose” button allows to select the classifier that you want to use.

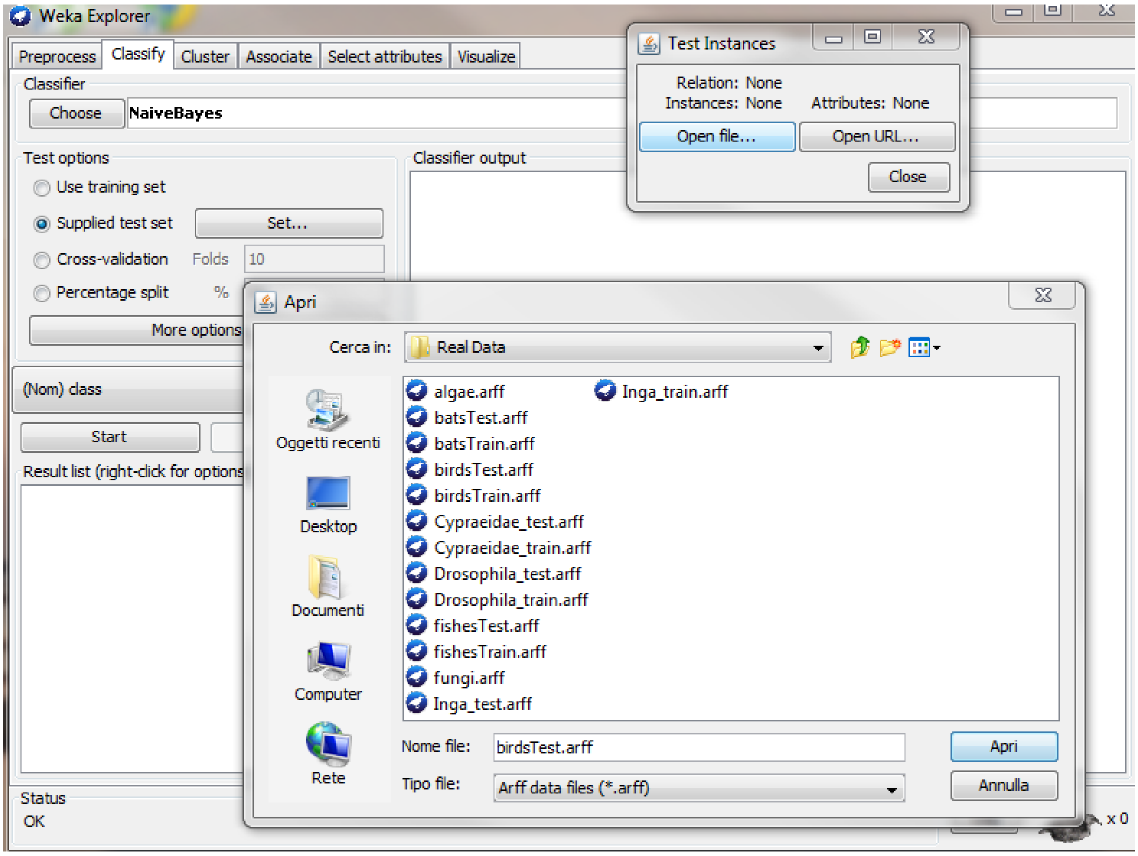
- Select the classifier family and type.



Status
OK

Classifier selection window

- Optionally, click to **Supplied test set** in order to select the dataset to be used as query set. Alternatively, you could use only one file, indicating the Percentage split or cross-validation.



Classifier: NaiveBayes

○ Use training set
● Supplied test set
○ Cross-validation Folds 10
○ Percentage split %

(Nom) class

Status: OK

Relation: None
Instances: None Attributes: None

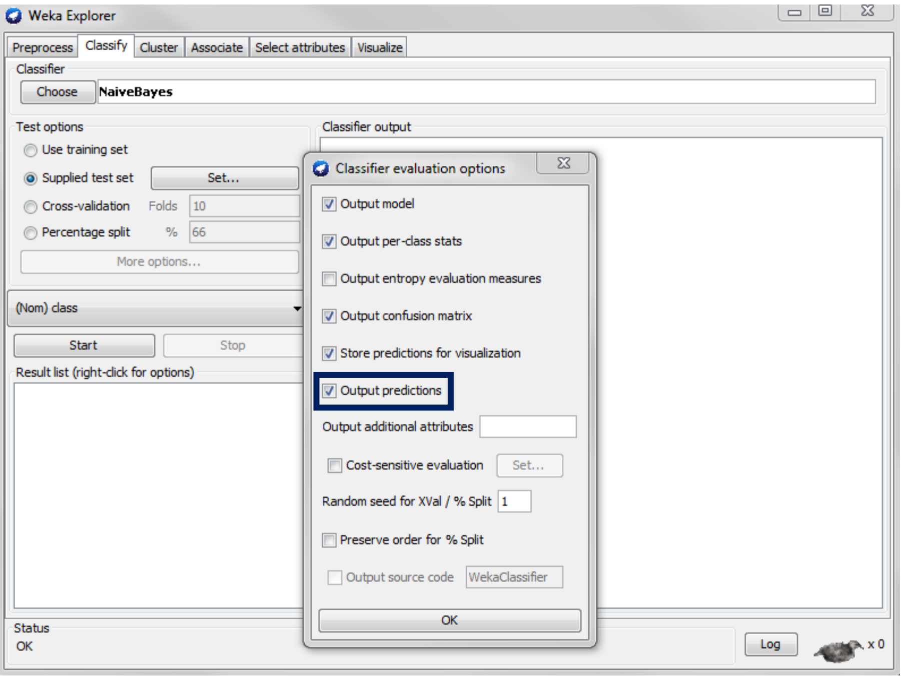
Cerca in: Real Data

algae.arff
batsTest.arff
batsTrain.arff
birdsTest.arff
birdsTrain.arff
Cypraeidae_test.arff
Cypraeidae_train.arff
Drosophila_test.arff
Drosophila_train.arff
fishesTest.arff
fishesTrain.arff
fungi.arff
Inga_test.arff
Inga_train.arff

Nome file: birdsTest.arff
Tipo file: Arff data files (*.arff)

Test set selection window

- If you have unknown barcodes sequences in the query set and you want the specimen to species assignments, through the **More options** button select the "Output prediction" option.



Classifier: NaiveBayes

○ Use training set
● Supplied test set
○ Cross-validation Folds 10
○ Percentage split % 66

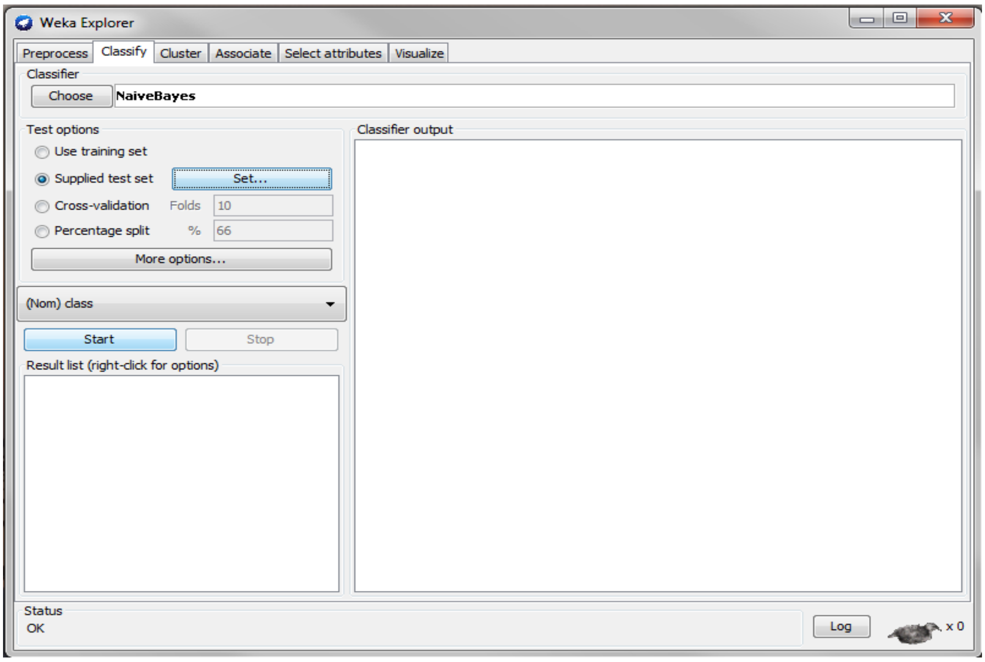
(Nom) class

Status: OK

☑ Output model
☑ Output per-class stats
□ Output entropy evaluation measures
☑ Output confusion matrix
☑ Store predictions for visualization
☑ Output predictions
Output additional attributes
□ Cost-sensitive evaluation
Random seed for XVal / % Split 1
□ Preserve order for % Split
□ Output source code WekaClassifier

Highlighting output prediction option

- Click the **Start** button to run the experiment.



Classifier: NaiveBayes

● Supplied test set

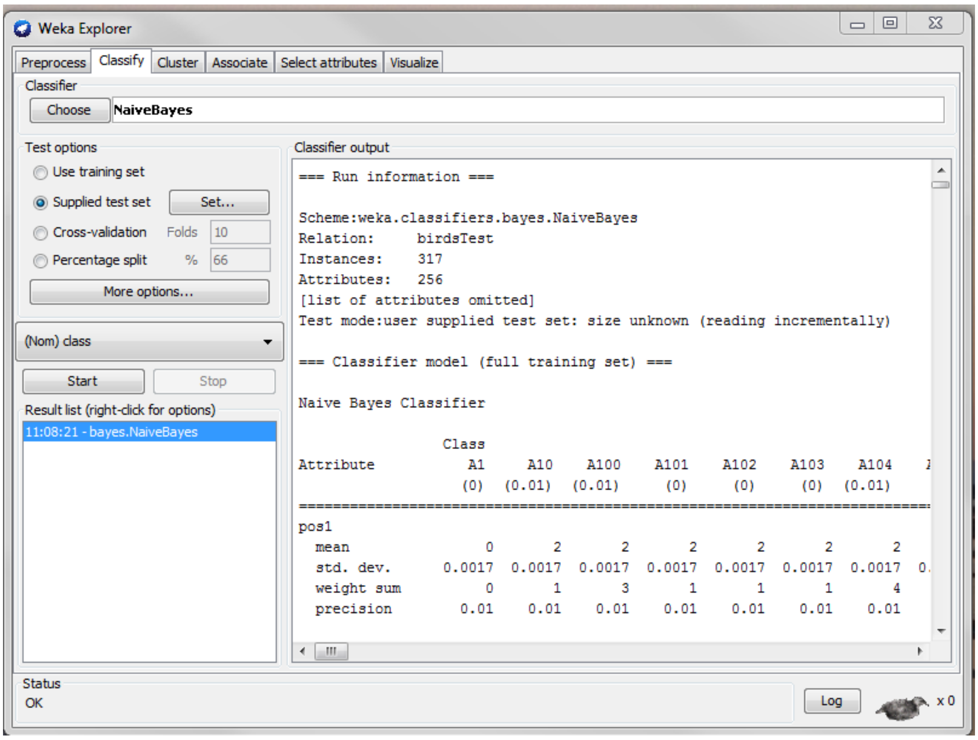
Cross-validation Folds 10

Percentage split % 66

(Nom) class

Status: OK

Start classification window



Classifier: NaiveBayes

● Supplied test set

Cross-validation Folds 10

Percentage split % 66

(Nom) class

Result list: 11:08:21 - bayes.NaiveBayes

```
=== Run information ===

Scheme:weka.classifiers.bayes.NaiveBayes
Relation:     birdsTest
Instances:    317
Attributes:   256
[list of attributes omitted]
Test mode:user supplied test set: size unknown (reading incrementally)

=== Classifier model (full training set) ===

Naive Bayes Classifier

                 Class
Attribute           A1      A10     A100     A101     A102     A103     A104    A
                   (0)   (0.01)   (0.01)      (0)      (0)      (0)   (0.01)
================================================================================
pos1
  mean               0        2        2        2        2        2        2
  std. dev.     0.0017   0.0017   0.0017   0.0017   0.0017   0.0017   0.0017   0.
  weight sum         0        1        3        1        1        1        4
  precision       0.01     0.01     0.01     0.01     0.01     0.01     0.01
```

Status: OK

Classifier output window

Data examples

The website http://dmb.iasi.cnr.it/supbarcodes.php contains the Fasta2Weka converter software package and some example DNA Barcode sequences data sets.

Contacts

Please contact Emanuel Weitschek (emanuel.weitschek@iasi.cnr.it) for comments and questions.