IA : Machine (Deep) Learning

Annexe : Weka

A. Zemmari

LaBRI - Université de Bordeaux

Jan. 09, 2018

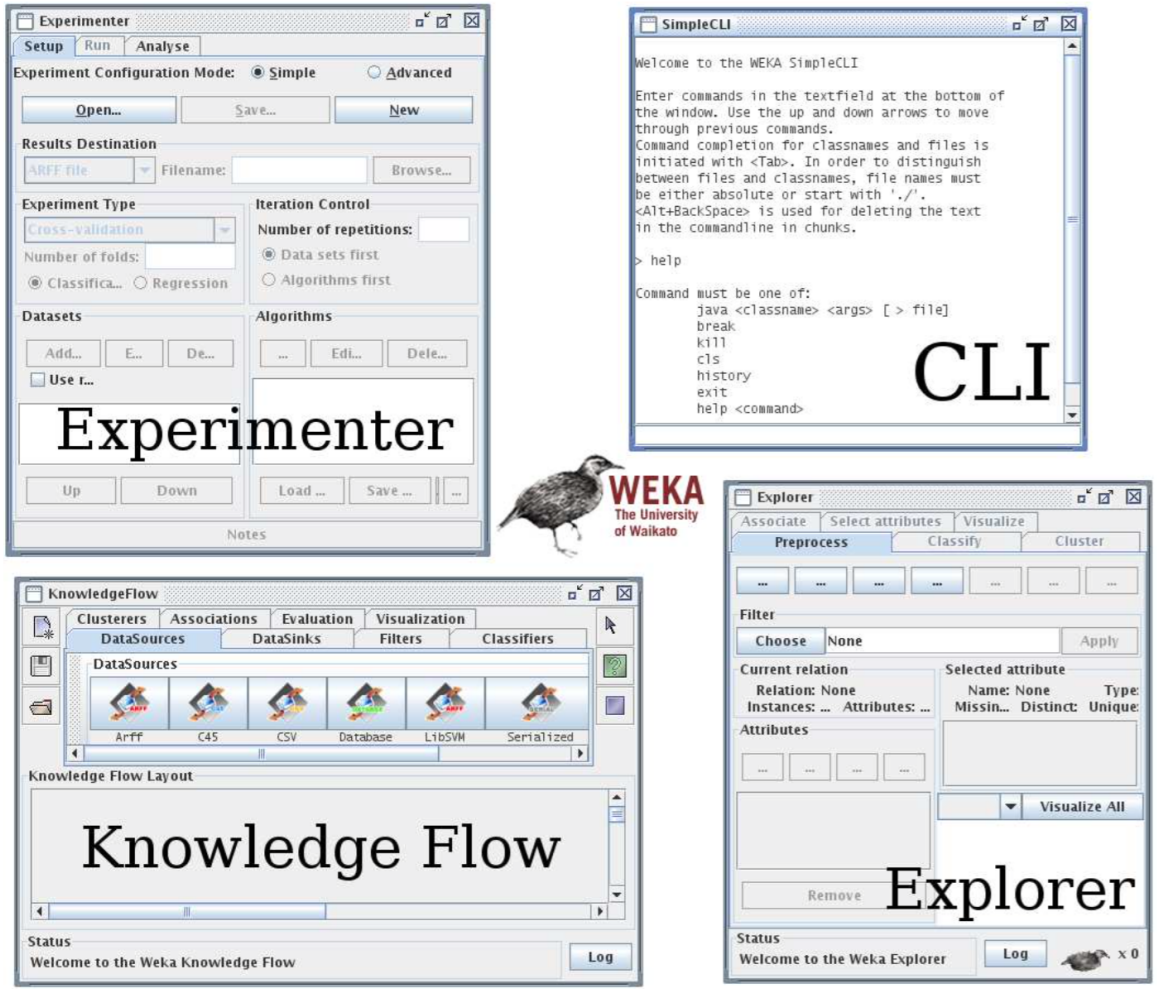
Présentation

- Weka (Waikato Environment for Knowledge Analysis) Environnement Waikato pour l'analyse de connaissances
- Suite de logiciels d'apprentissage automatique et d'exploration de données écrite en Java
- Développée à l'université de Waikato en Nouvelle-Zélande
- Disponible pour toutes les plateformes : Windows, Mac OS X, Linux etc.
- Documentation riche et communauté large
  - Le livre Data Mining : Practical Machine Learning Tools and Techniques (troisième édition)
  - API : http://weka.sourceforge.net/doc.stable/
  - Wiki : http://weka.wikispaces.com/
  - FAQ : http://weka.wikispaces.com/FAQ

L'outil Weka

- Outils de pré-traitement des données (filtering)
  - Sélection, transformation, combinaison d'attributs, normalisation, re-échantillonnage, etc.
- Algorithmes pour l'exploration de données
  - Clustering, classification, régression, etc.
- Analyse des résultats
  - Evaluation de performances, comparaison d'algorithmes, etc.
- Plusieurs interfaces
  - Graphiques (Explorer, Experimenter et Knowledge Flow)
  - En ligne de commande (CLI)

Aperçu de l'interface de Weka



Traitement des données

Format de données :

- Le format d'entrée par défaut de Weka est le ARFF (Attribute Relation File Format)
- D'autres formats peuvent être importés : CSV, binaire, BDD SQL (avec JDBC), à partir d'une URL, etc.

Traitement des données

- Caractéristiques du format de fichier ARFF :
  - Les commentaires sont précédés de %
  - Définition du nom de l'ensemble de données avec `@relation`
  - Le nom doit être aussi compréhensible que possible
    `@relation temperatureMaison_14jours`
  - Définition des features avec `@attribute`
    - Attributs nominaux suivis des valeurs entre accolades
      `@attribute outlook {sunny, overcast, rainy}`
    - Attributs numériques avec `real`
      `@attribute temperature real`
    - Attributs chaines avec `string`, les valeurs doivent être entre doubles guillemets
      `@attribute unTexte string`
    - Attributs dates avec `date` (`yyyy-MM-dd-THH :mm :ss`)
      `@attribute uneDate date`
  - `@data` signale le début des instances

Traitement des données

Exemple d'un fichier ARFF :

```
% Ensemble de donnees sur la meteo
@relation weather
% Definition des features
@attribute outlook {sunny, overcast, rainy}
@attribute temperature real
@attribute humidity real
@attribute windy {TRUE, FALSE}
@attribute play {yes, no}
% Debut des instances
@data
sunny,85,85,FALSE,no
sunny,80,90,TRUE,no
overcast,83,86,FALSE,yes
rainy,70,96,FALSE,yes
...
```

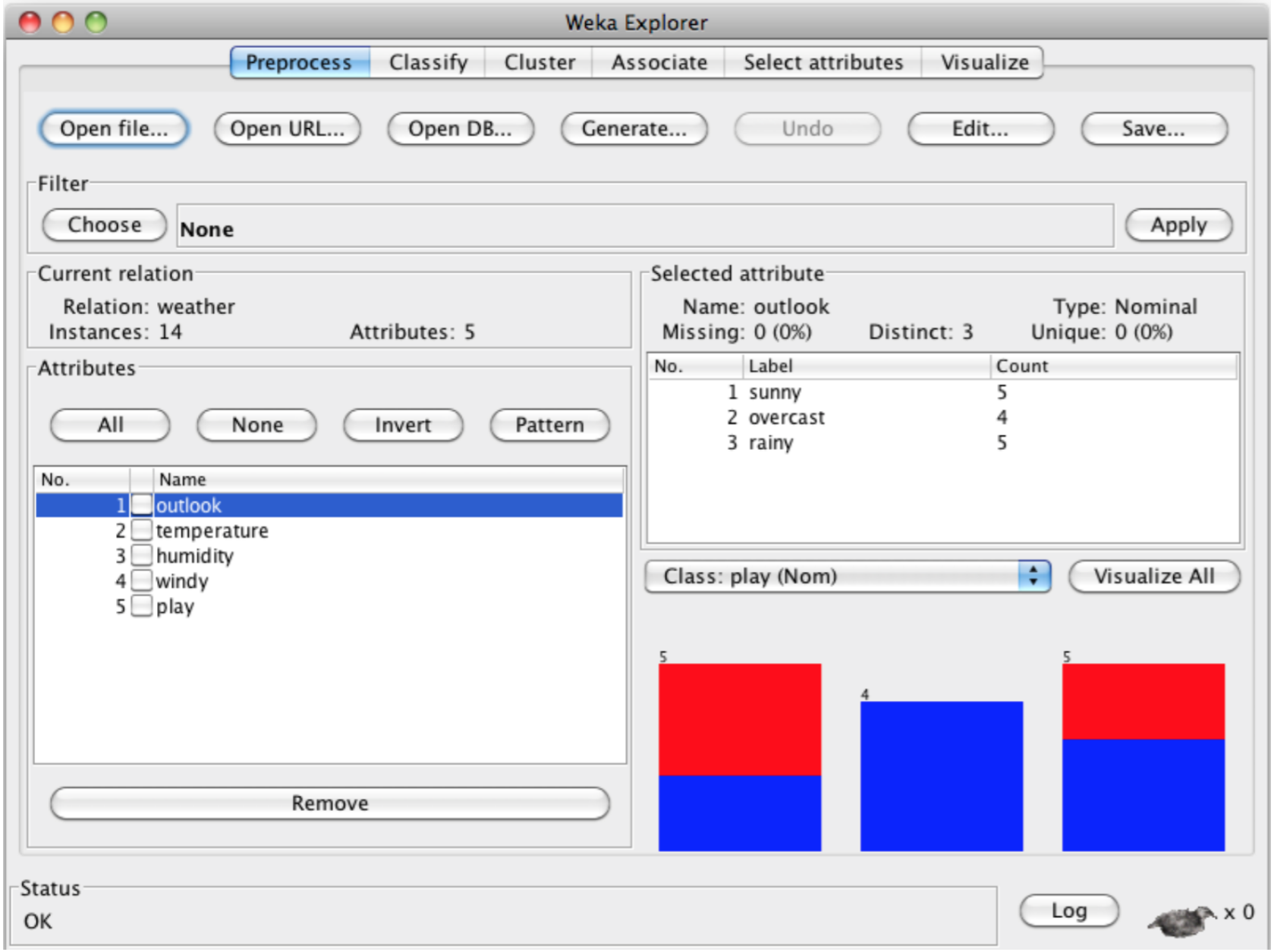
Pré-traitement des données (1)

- Les pré-traitements dans Weka sont effectués grâce aux filtres
- Les filtres permettent de modifier les ensemble de données : supprimer ou ajouter des attributs, ré-échantillonner, supprimer des exemples, etc.
- Le package `weka.filters` dispose de deux types de filtres
  - Filtres non supervisés
  - Filtres supervisés, i.e. tirant parti de l?information de la classe
- Il existe des filtres pour les attributs et pour les exemples

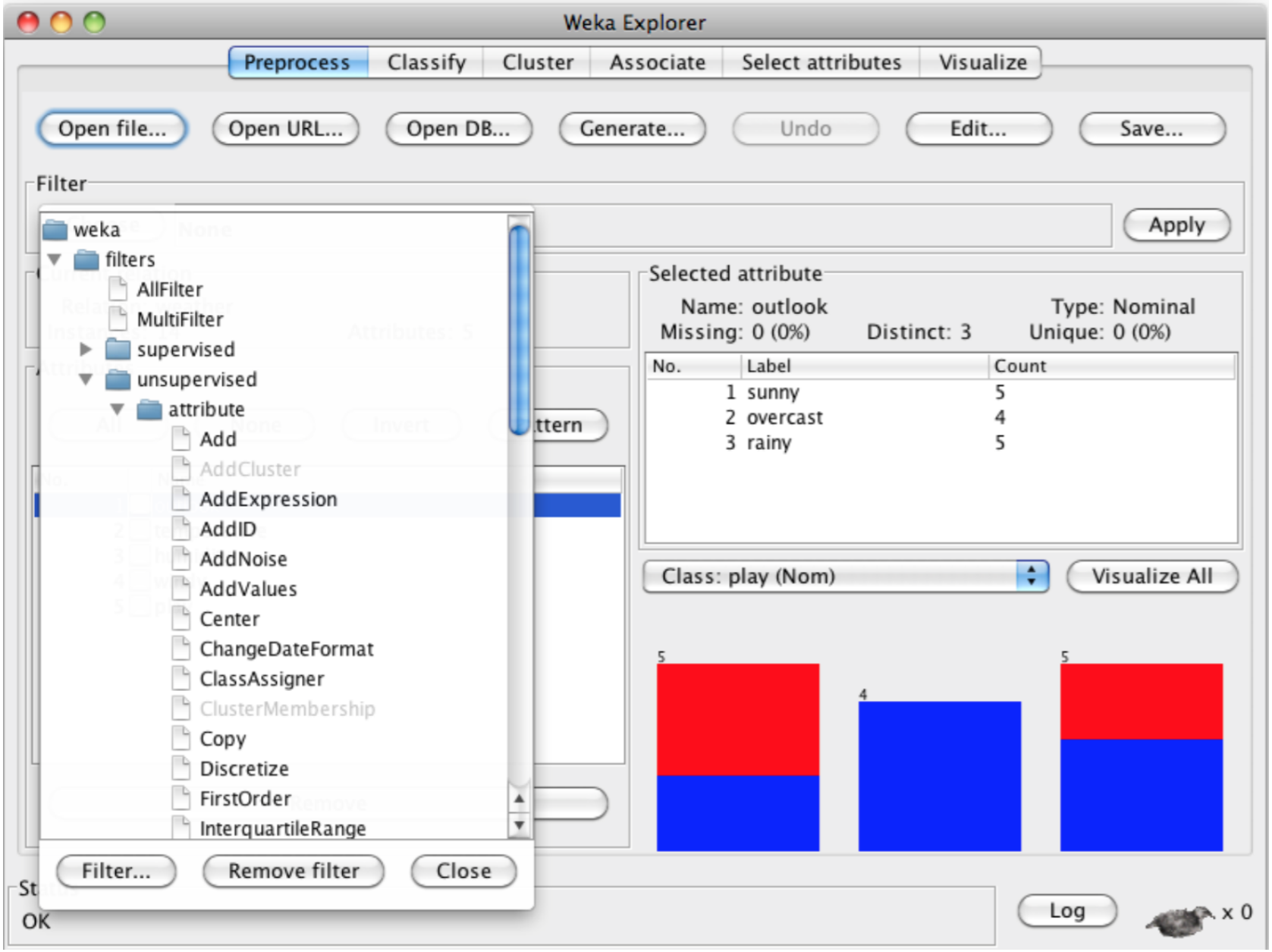
Pré-traitement des données (2)

- Plusieurs dizaines de filtres
  - Ajout, suppression et copie d?attributs (`Add`, `Remove` et `Copy`)
  - Ajout de bruit (`AddNoise`)
  - Fusion d'attributs (`Merge`)
  - Conversion (`NominalToBinary`, `StringToBinary`, etc.)
  - Anonymiser (`Obfuscate`)
  - Méthode de discrétisation (`Discretize`)

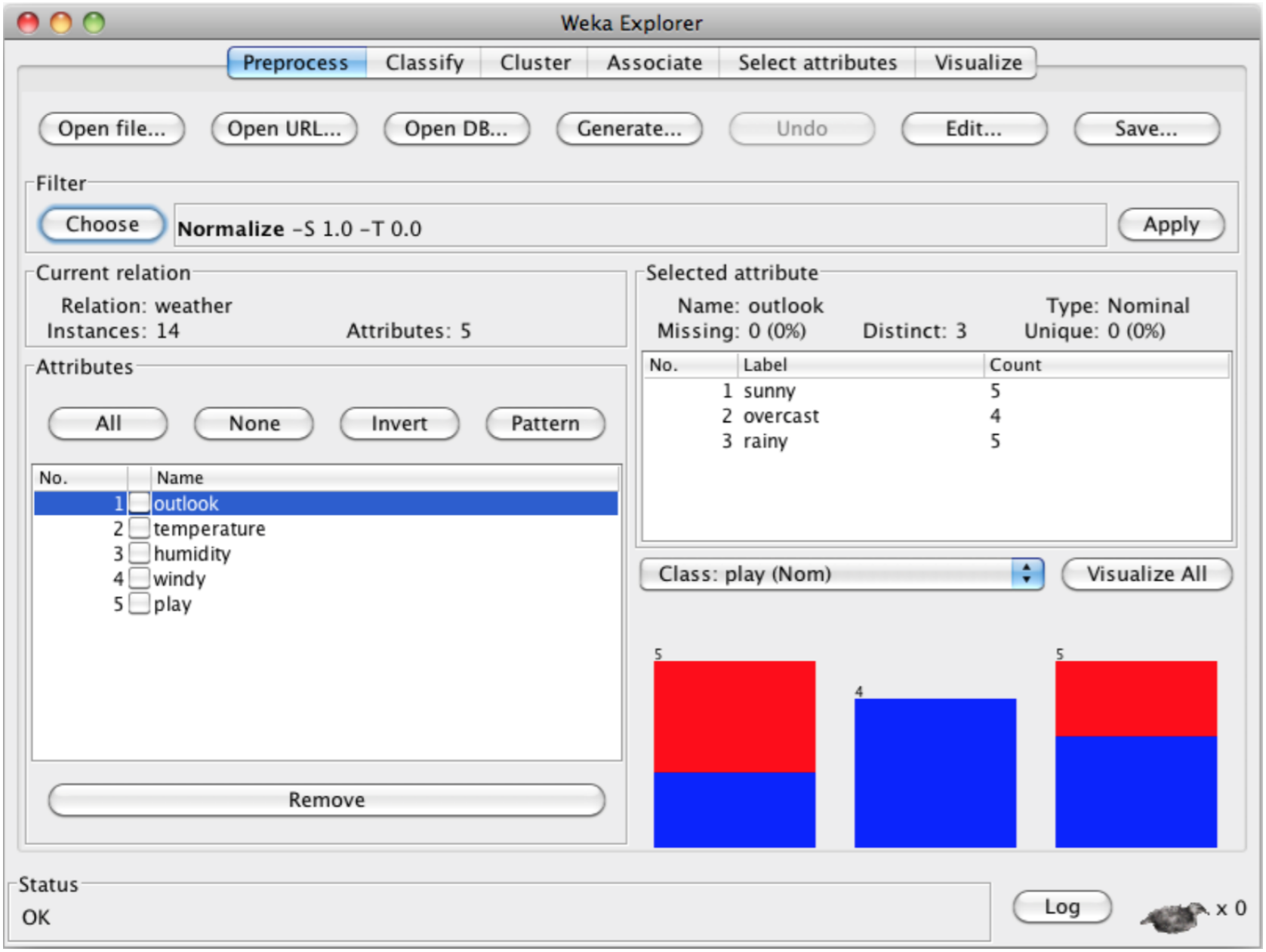
Exemple de filtre : normalisation



Exemple de filtre : normalisation



Exemple de filtre : normalisation



Exemple de filtre : normalisation

Viewer

Relation: weather

No.	outlook Nominal	temperature Numeric	humidity Numeric	windy Nominal	play Nominal
1	sunny	85.0	85.0	FALSE	no
10	rainy	75.0	80.0	FALSE	yes
11	sunny	75.0	70.0	TRUE	yes
12	overcast	72.0	90.0	TRUE	yes
13	overcast	81.0	75.0	FALSE	yes
14	rainy	71.0	91.0	TRUE	no
2	sunny	80.0	90.0	TRUE	no
3	overcast	83.0	86.0	FALSE	yes
4	rainy	70.0	96.0	FALSE	yes
5	rainy	68.0	80.0	FALSE	yes
6	rainy	65.0	70.0	TRUE	no
7	overcast	64.0	65.0	TRUE	yes
8	sunny	72.0	95.0	FALSE	no
9	sunny	69.0	70.0	FALSE	yes

Undo OK Cancel

Viewer

Relation: weather-weka.filters.unsupervised.attrib...

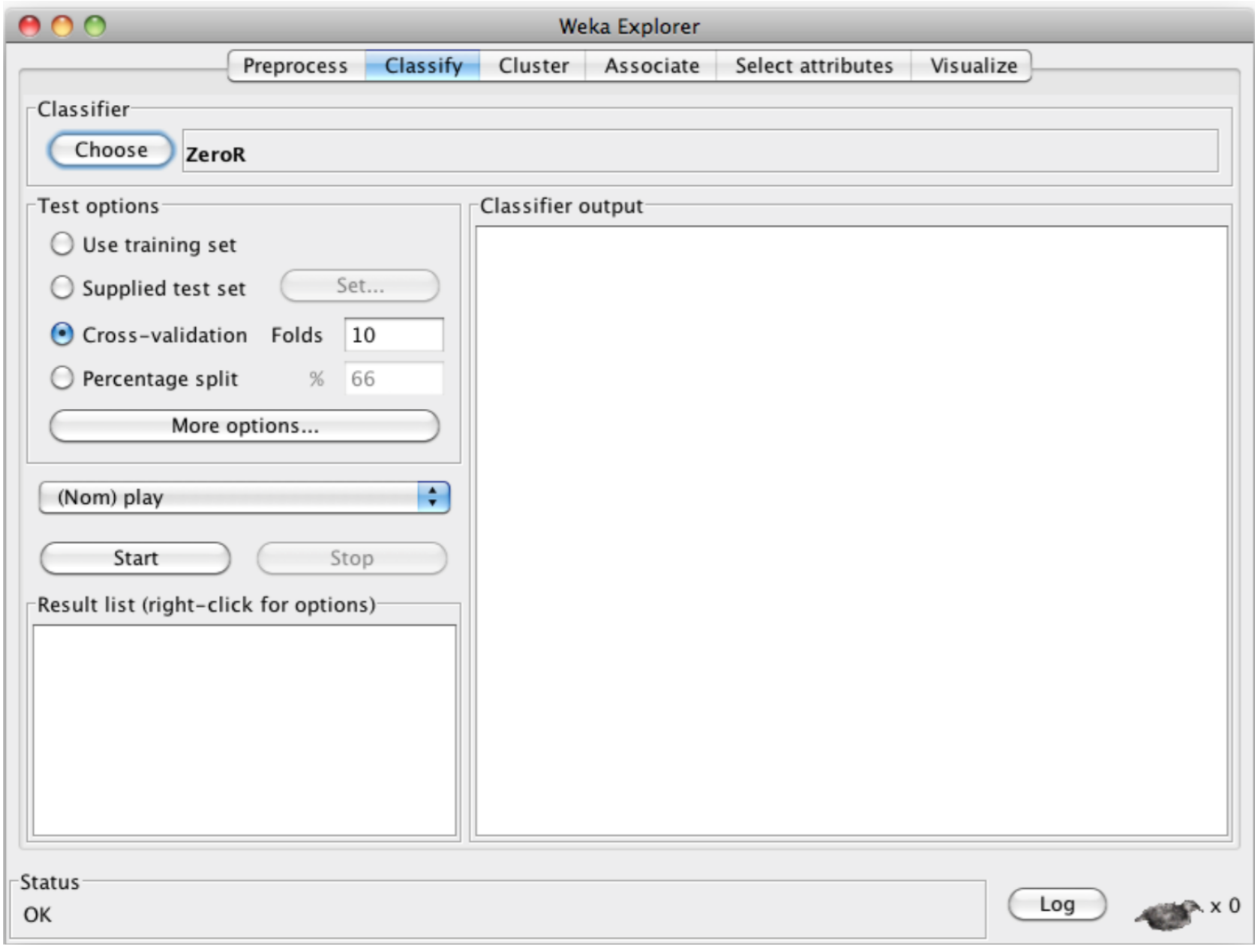
No.	outlook Nominal	temperature Numeric	humidity Numeric	windy Nominal	play Nominal
1	sunny	1.0	0.6451...	FALSE	no
10	rainy	0.5238095...	0.4838...	FALSE	yes
11	sunny	0.5238095...	0.1612...	TRUE	yes
12	overcast	0.3809523...	0.8064...	TRUE	yes
13	overcast	0.8095238...	0.3225...	FALSE	yes
14	rainy	0.3333333...	0.8387...	TRUE	no
2	sunny	0.7619047...	0.8064...	TRUE	no
3	overcast	0.9047619...	0.6774...	FALSE	yes
4	rainy	0.2857142...	1.0	FALSE	yes
5	rainy	0.1904761...	0.4838...	FALSE	yes
6	rainy	0.0476190...	0.1612...	TRUE	no
7	overcast	0.0	0.0	TRUE	yes
8	sunny	0.3809523...	0.9677...	FALSE	no
9	sunny	0.2380952...	0.1612...	FALSE	yes

Undo OK Cancel

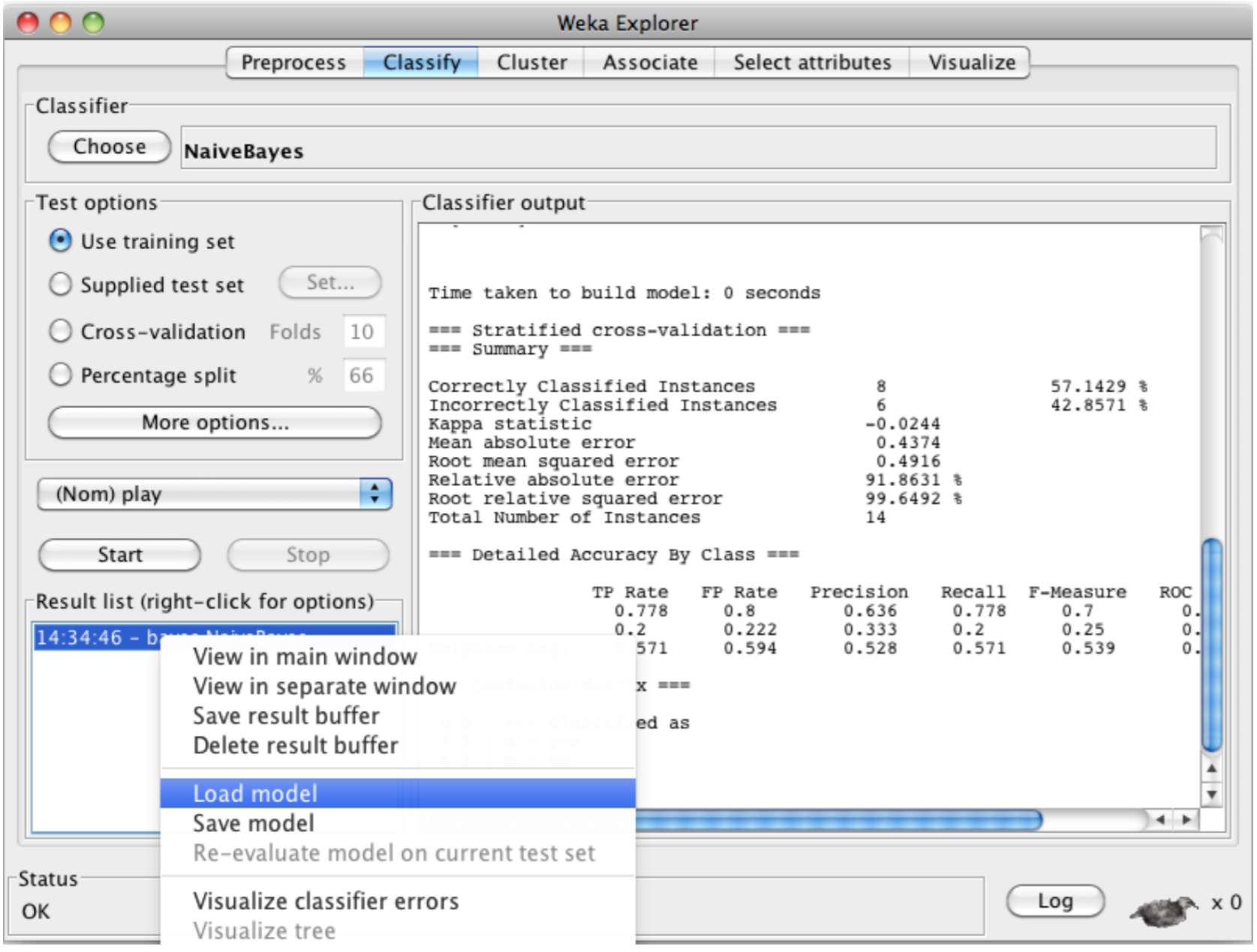
Contenu du Toolkit

- 49 outils pour du data preprocessing
- 76 algorithmes de classification/regression
- 8 algorithmes de clustering
- 3 algorithmes pour trouver les règles dassociation
- ...

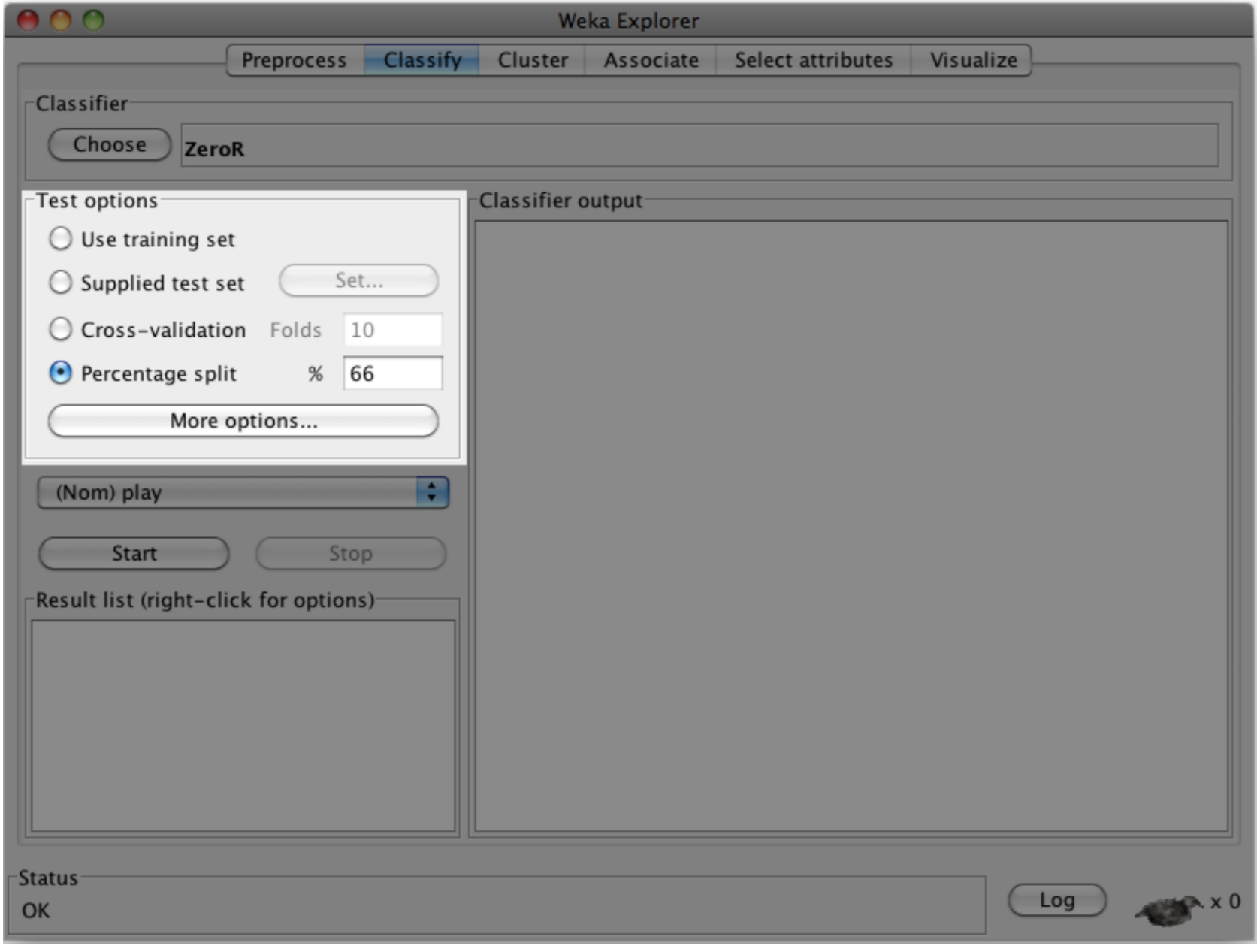
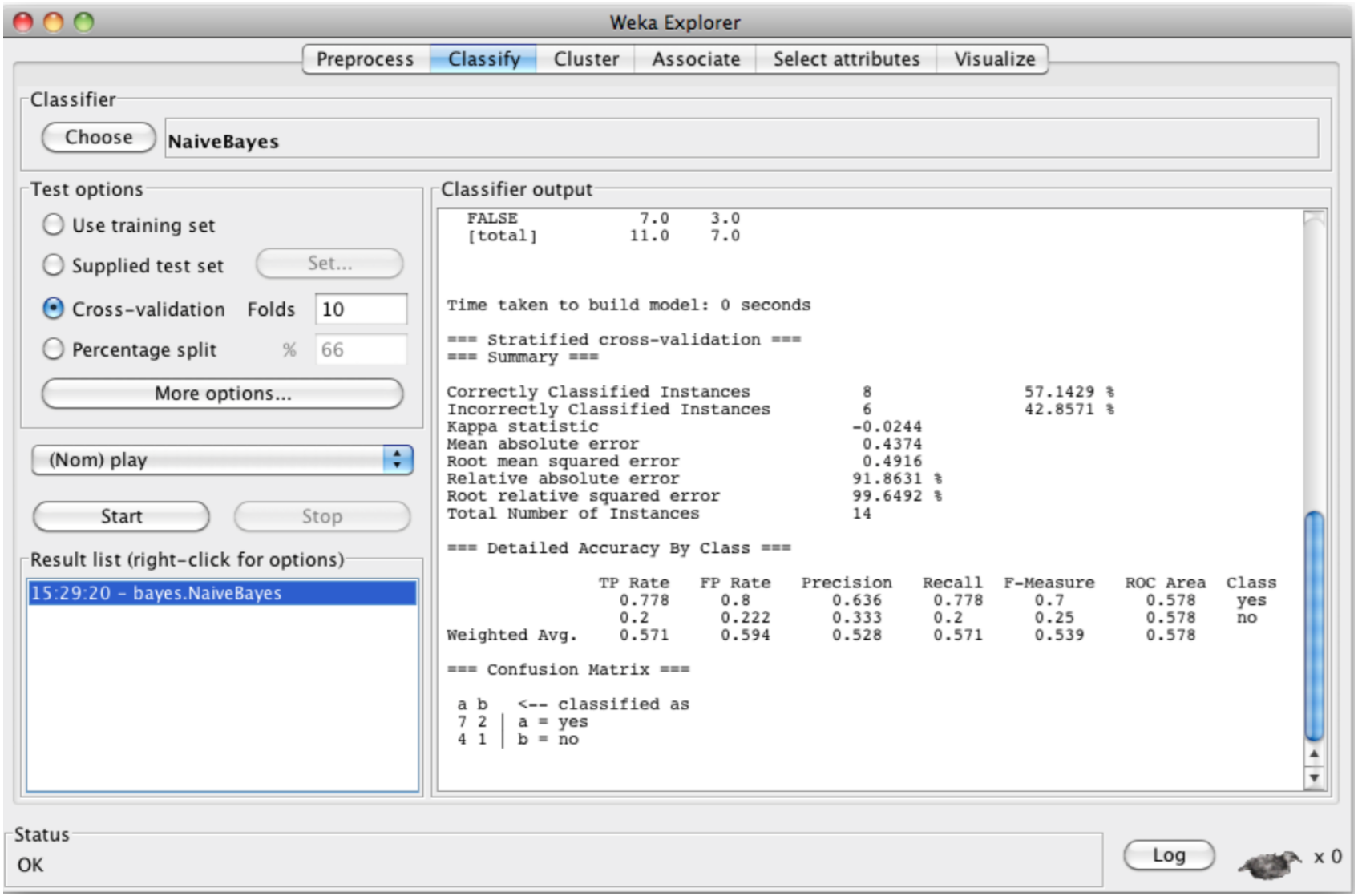
L'onglet classify



Entraînement et sauvegarde des modèles



Découpage des données



Evaluation du modèle