Approche Data Mining par WEKA

WEKA, un logiciel libre d'apprentissage et de data mining

Yves Lechevallier

INRIA-Rocquencourt

E_mail : Yves.Lechevallier@inria.fr

WEKA 3.4 : Plan

- Présentation de WEKA 3.4
- Format ARFF
- WEKA Explorer
- WEKA Experiment Environment
- WEKA KnowledgeFlow

WEKA 3.4

THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato

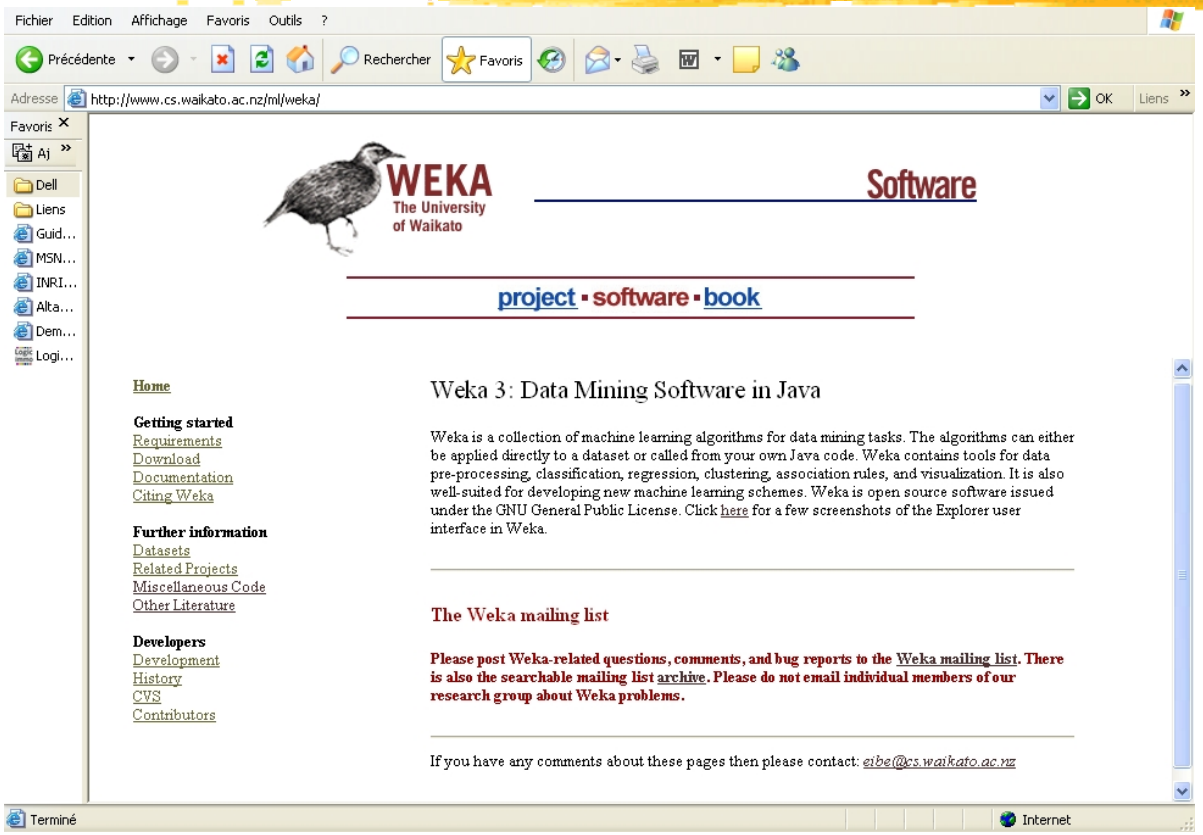
Site WEB :

http://www.cs.waikato.ac.nz/ml/weka/

Weka Machine Learning Project (GNU Public License)

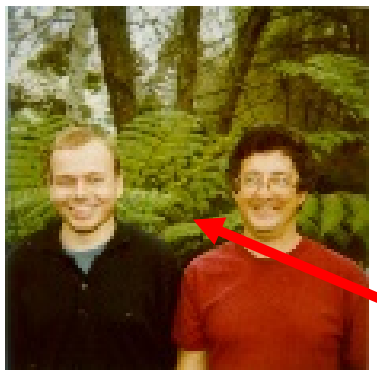
- Format ARFF
- Beaucoup de méthodes de classification supervisées
- Quelques méthodes de classification automatiques
- Quelques méthodes de recherches de sous-ensembles fréquents
- Fichiers de résultats standardisés

WEKA 3.4 Le Site



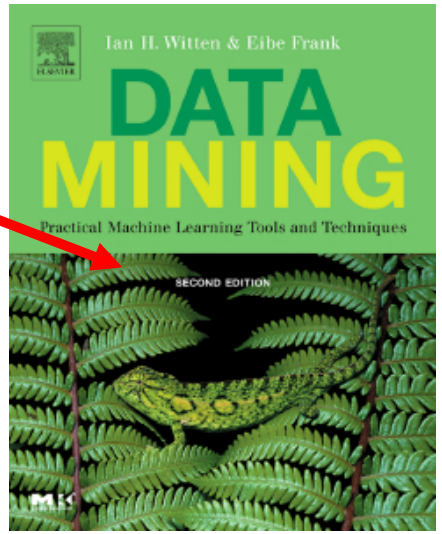
WEKA 3.4 Documentation (1/2)

Ian H. Witten and Eibe Frank (2005) "*Data Mining: Practical machine learning tools and techniques*", 2nd Edition, Morgan Kaufmann, San Francisco, 2005.



Le livre

Les auteurs



WEKA 3.4 Documentation (2/2)

Documentation en ligne

A presentation demonstrating all graphical user interfaces in Weka. (Warning: this is a large Powerpoint file.)

Une présentation très bien faite. A lire avant de commencer.

An introduction , written by Alex K. Seewald, to using Weka 3.4.4 from the command line.

A lire pour les concepts de base

A page documenting the ARFF data format used by Weka.

A lire car c'est le format d'entrèe de Weka

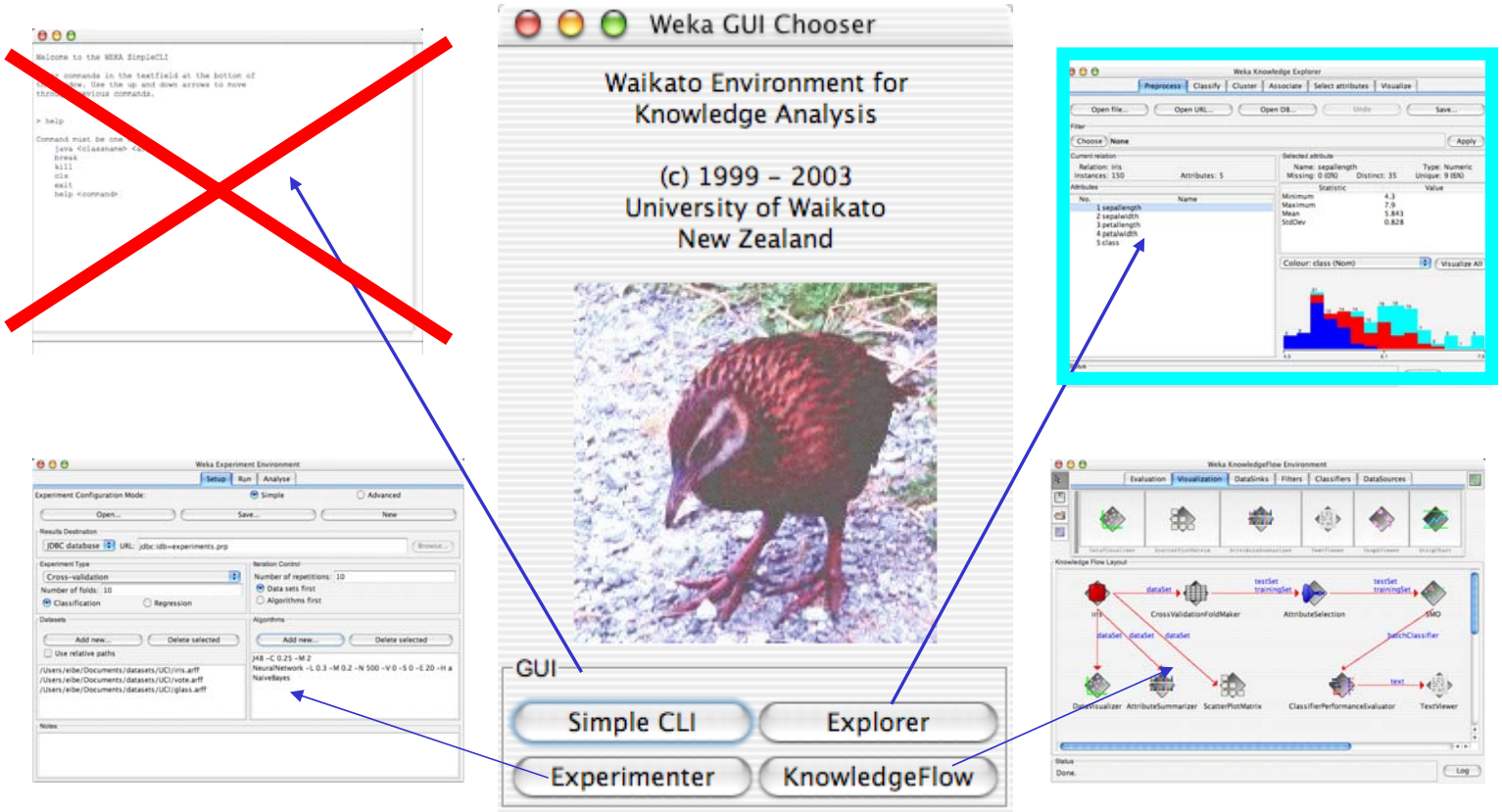
A page describing how to load your MS Access database into the Weka Explorer.

Indispensable pour le monde Windows

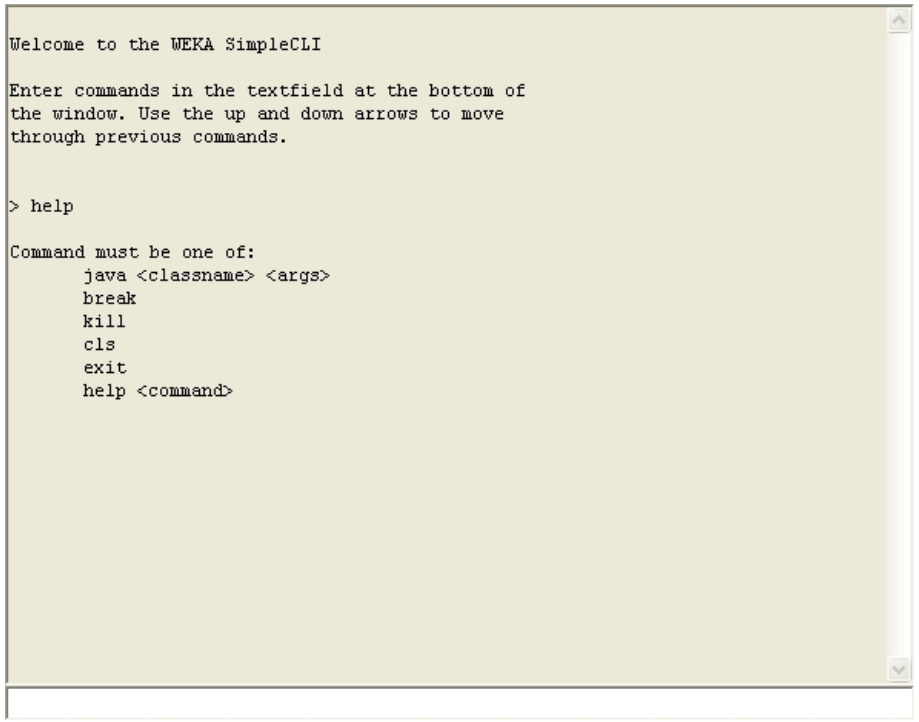
A page describing the use of XML in WEKA.

pour le futur

GUI : Interfaces utilisateurs



Interface «ligne de commande »



Les concepts

- Un tableau des données ou une collection d'exemples (dataset).
- Chaque ligne de ce tableau représente une observation qui est décrite par un vecteur (instance)
- Chaque colonne représente une variable (attribute) qui peut être quantitative (numeric), qualitative (nominal) ou textuelle (string).

Format ARFF, un exemple

```
% 1. Title: Iris Plants Database
%      (a) Creator: R.A. Fisher
%      (b) Donor: Michael Marshall (MARSHALL%PLU@io.arc.nasa.gov)
%      (c) Date: July, 1988
@RELATION iris

 @ATTRIBUTE sepallength  NUMERIC
 @ATTRIBUTE sepalwidth   NUMERIC
 @ATTRIBUTE petallength  NUMERIC
 @ATTRIBUTE petalwidth   NUMERIC
 @ATTRIBUTE class        {Iris-setosa,Iris-versicolor,Iris-virginica}
```

Entête

```
@DATA
  5.1,3.5,1.4,0.2,Iris-setosa
  4.9,3.0,1.4,0.2,Iris-setosa
  4.7,3.2,1.3,0.2,Iris-setosa
  4.6,3.1,1.5,0.2,Iris-setosa
  5.0,3.6,1.4,0.2,Iris-setosa
  5.4,3.9,1.7,0.4,Iris-setosa
  4.6,3.4,1.4,0.3,Iris-setosa
```

Les données

Format ARFF, l'entête

Le nom du tableau de données

@RELATION iris

@ATTRIBUTE sepallength NUMERIC
@ATTRIBUTE sepalwidth NUMERIC
@ATTRIBUTE petallength NUMERIC
@ATTRIBUTE petalwidth NUMERIC

Variables continues

@ATTRIBUTE class {Iris-setosa,Iris-versicolor,Iris-virginica}

Variable discrète

Format ARFF, Les données

@DATA

5.1,3.5,1.4,0.2,	Iris-setosa
4.9,3.0,1.4,0.2,	Iris-setosa
4.7,3.2,1.3,0.2,	Iris-setosa
4.6,3.1,1.5,0.2,	Iris-setosa
5.0,3.6,1.4,0.2,	Iris-setosa
5.4,3.9,1.7,0.4,	Iris-setosa
4.6,3.4,1.4,0.3,	Iris-setosa

Numérique Nominal

Le séparateur est la virgule.

Si le nom d'une modalité contient un blanc mettre entre guillemets

Explorer : Interface « utilisation des méthodes »

Objectifs :

- Le Weka Explorer permet de lancer une méthode à partir d'un fichier ARFF.
- Les résultats sont mis sous la forme d'un fichier texte normalisé.
- Permet de sélectionner la méthode la mieux adaptée ou la plus efficace.

Explorer : Multiplot

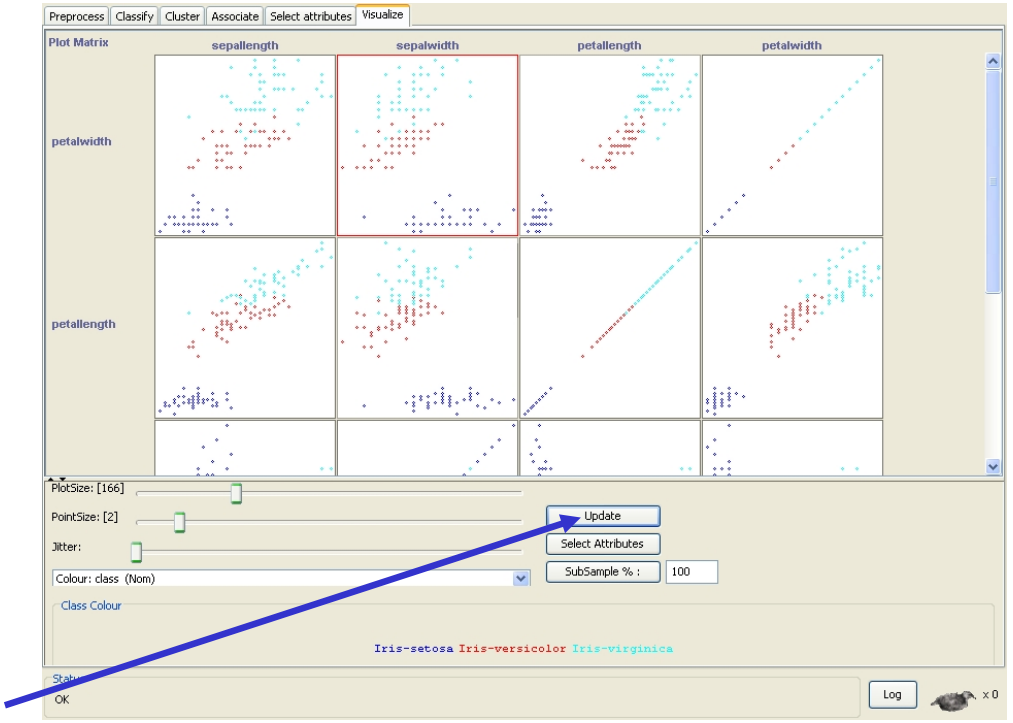
Sélection de

• la taille de l'image,

•des variables

•de la grosseur des points

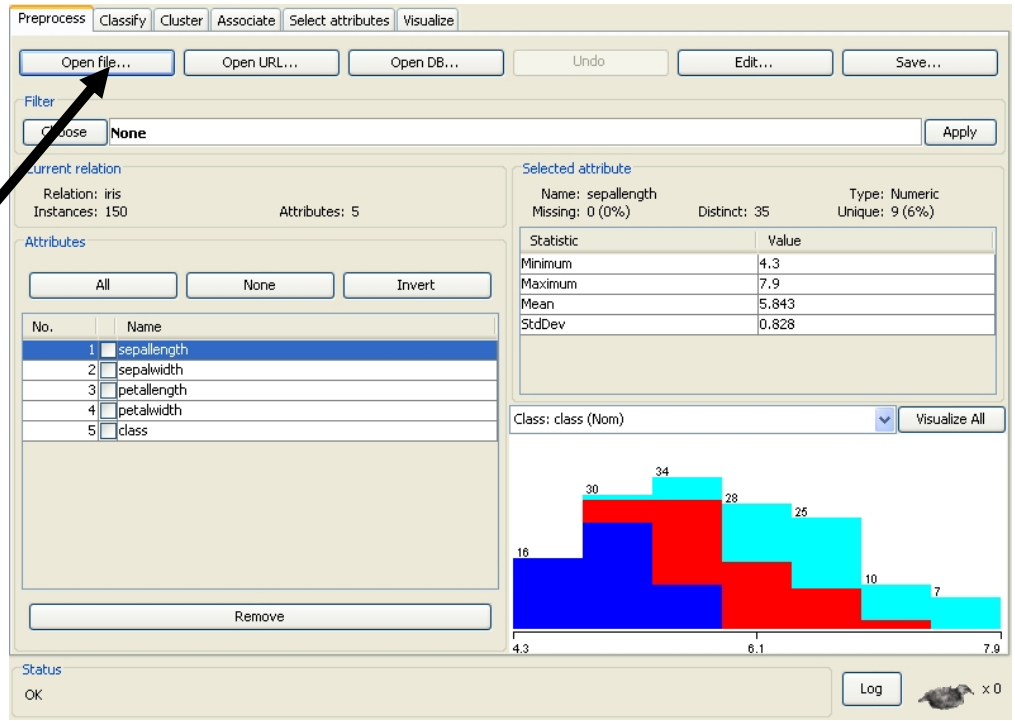
Ne pas oublier



Explorer : Interface « utilisation des méthodes »

(1)

Lecture du tableau de données au format ARFF



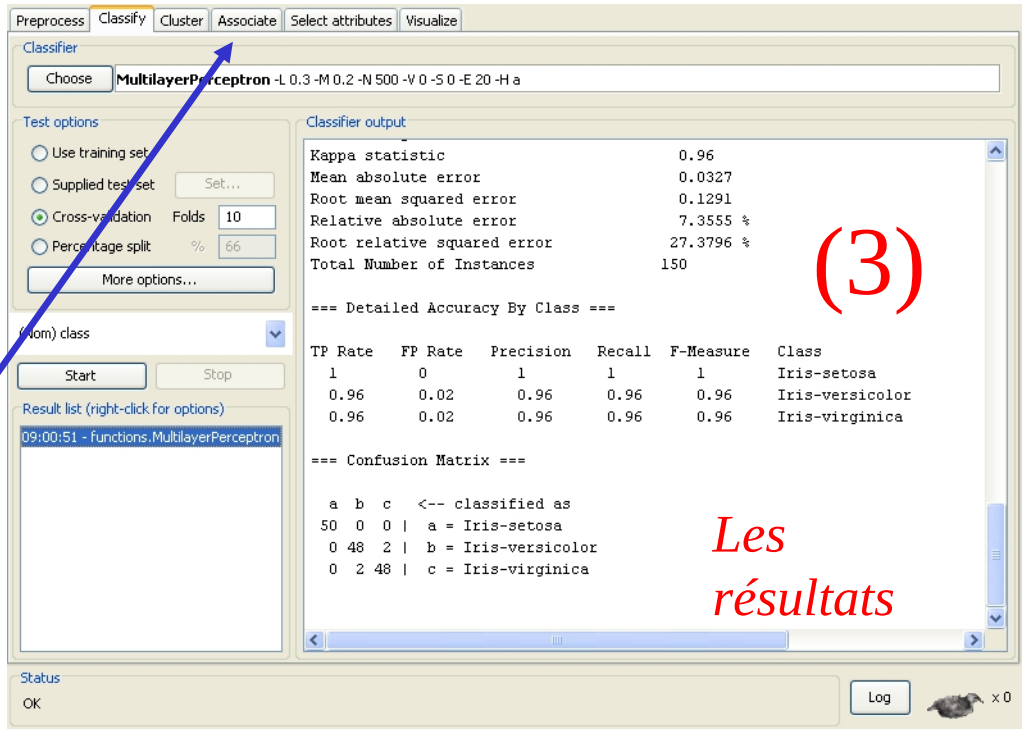
La dernière variable est la variable à prédire.

Explorer : Interface « utilisation des méthodes »

(2)

Choix de la méthode

- Visualize
- Select attribute
- Classify
- Cluster
- Associate



(3)

Les résultats

Règle de Bayes d'erreur minimale

$$\forall x \quad Y^*(x) = k \text{ où k est telque } \Pr(k/x) = \max \Pr(h/x)$$

Cette définition est peu opérationnelle, en effet, on connaît rarement la probabilité d'un classement sachant une description.

Théorème de Bayes $\quad \Pr(k/x) = \dfrac{\pi_k L_k(x)}{L(x)}$

$$\pi_k = \Pr[Y = k]$$

$$L_k(x) = \Pr[X = x / Y = k] \text{ est la densité de la classe } k$$

$$\forall x \quad Y^*(x) = k \text{ où } k \text{ est telque } \Pr(k/x) = \max \pi_k L_k(x)$$

Les descriptions suivent une loi normale

Le descripteur X des exemples est constitué de p descripteurs numériques et que sa distribution, conditionnellement aux classes, suit une loi normale multidimensionnelle centrée sur le vecteur μ_k et de matrice de variance-covariance Σ_k .

La vraisemblance conditionnelle de X pour la classe k s'écrit alors

$$L_k(x) = \left((2\pi)^p \det \Sigma_k\right)^{-\frac{1}{2}} \exp\left(-\tfrac{1}{2}(x-\mu_k)^t \Sigma_k^{-1}(x-\mu_k)\right)$$

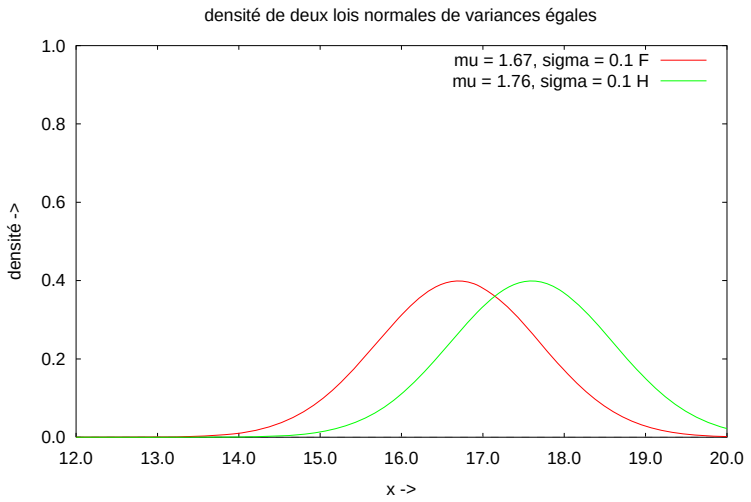
Exemple

Les variances et les probabilités a priori sont égales

La taille moyenne des femmes est égale à 1,67

La taille moyenne des hommes est égale à 1,76

μ_1=1,67 et μ_2=1,76

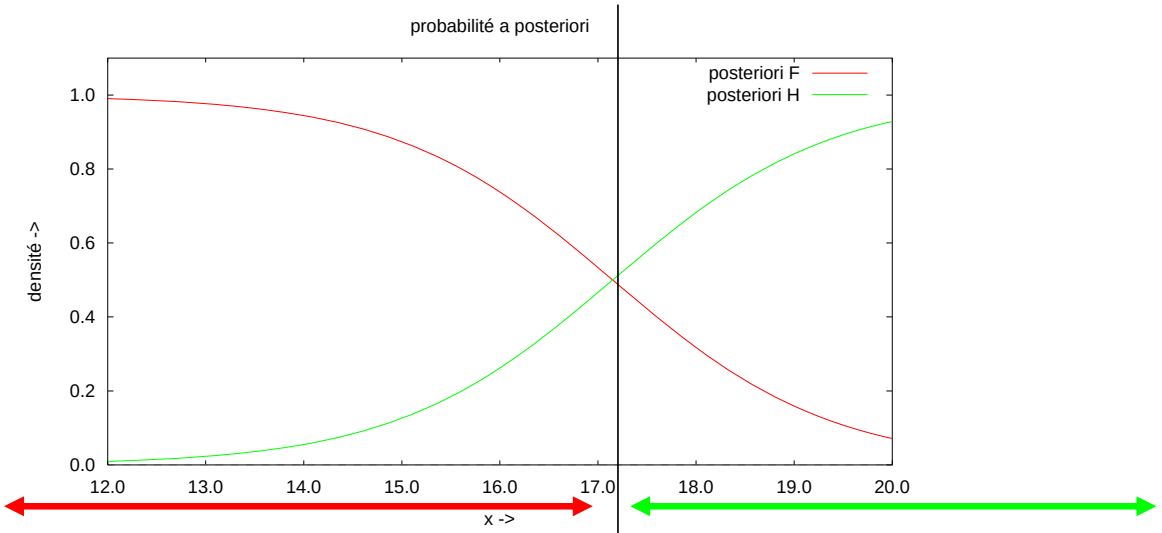


$L_k(x)$

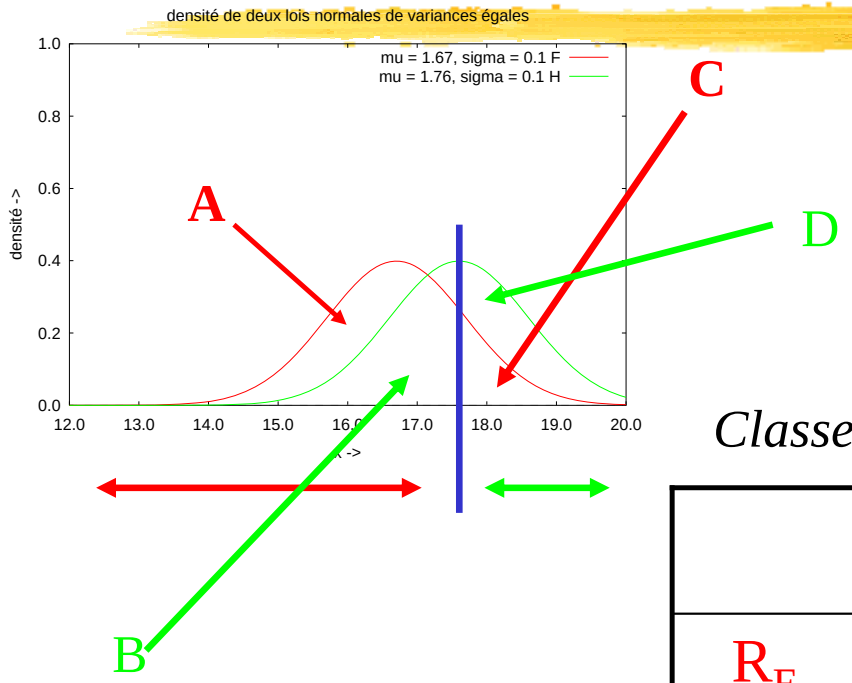
Règle de Bayes

$$\Pr(k / x) = \frac{\pi_k L_k(x)}{L(x)}$$

Cette règle minimise le pourcentage de mauvais classement



Construction d'un tableau de confusion à partir d'une fonction de décision



Qualité de la décision :

(A+D)/(A+B+C+D)

Classes a priori

	F	H
R_F	A	B
R_H	C	D

Classes d'affectation

Qualité d'un score

- Chaque sortie du réseau est associée à une classe a priori.
- L'objectif est d'analyser les scores de cette sortie
- Les exemples sont les observations de la classe a priori associée à cette sortie
- Les contre-exemples sont les observations des autres classes

courbe ROC (1/3)
Receiver Operating Characteristic curve

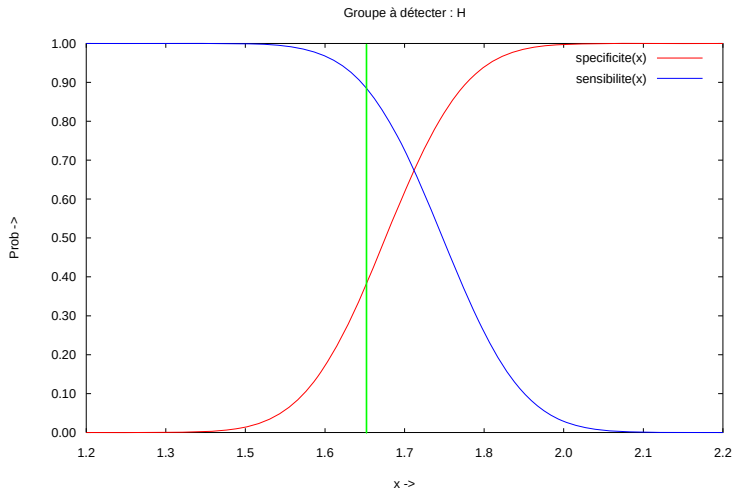
Pour un score *s* nous avons quatre comptages

- (A) Les Vrais Positifs sont les exemples ayant une valeur supérieure à *s*.
- (D) Les Vrais Négatifs sont les contre-exemples ayant une valeur inférieure à *s*.
- (B) Les Faux Négatifs sont les exemples ayant une valeur inférieure à *s*.
- (C) Les Faux Positifs sont les contre-exemples ayant une valeur supérieure à *s*.

Courbe ROC (2/3)

- On se fixe la classe a priori G et F est l'ensemble des autres classes a priori
- La sensibilité du score s est égale à P[S>s/G], la sensibilité est le pourcentage de Vrais Positifs
- La spécificité du score s est égale à P[S<s/F], la spécificité est le pourcentage de Vrais Négatifs

Courbe ROC



Si s=1,6 on a 90% des exemples dépassent cette valeur et 40% des contre-exemples sont en dessous de cette valeur

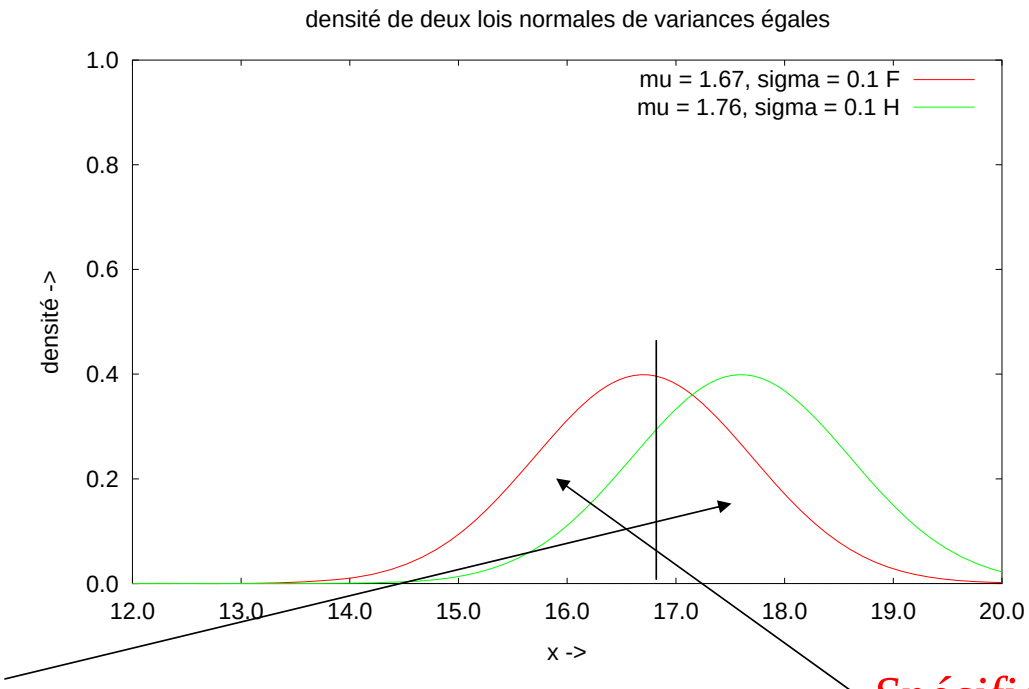
Quand le score augmente

la sensibilité diminue cela signifie que le % d'exemples dépassant cette valeur diminue: A/(A+C)

La spécificité augmente cela signifie que le % de contre-exemples en dessous de cette valeur augmente: D/(B+D)

Courbe ROC

$L_k(x)$



Sensibilité=*VP*

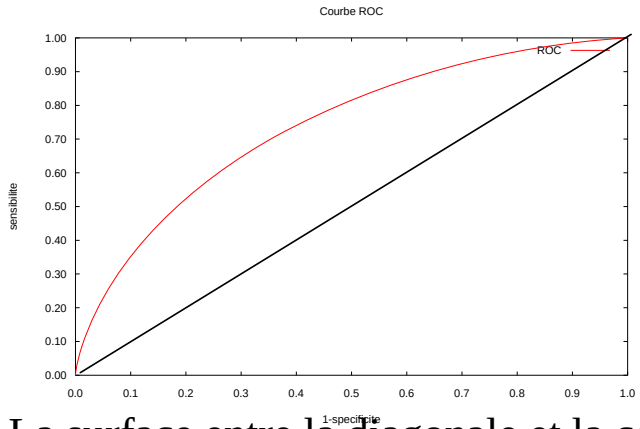
Spécificité=*VN*

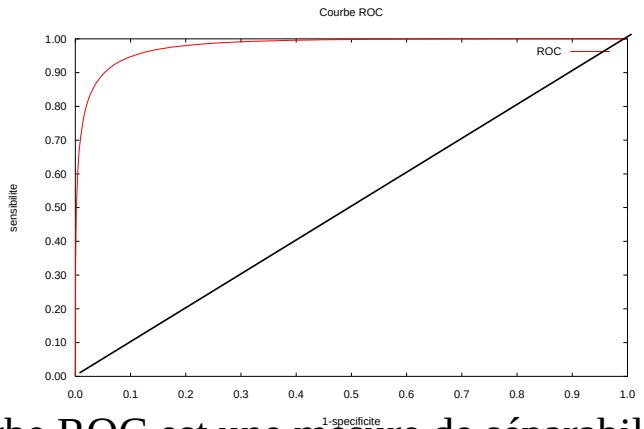
Courbe ROC : interprétation

La diagonale représente la courbe ROC d'un échantillon d'exemples et contre-exemples complètement mélangés

La courbe ROC de gauche est celle de notre exemple (μ_1=1,67 et μ_2=1,76)

La courbe ROC de droite est celle obtenue avec μ1=1,57 et μ_2=1,86





La surface entre la diagonale et la courbe ROC est une mesure de séparabilité des exemples avec les contre-exemples.

F-mesure : évaluation externe

L'évaluation de la qualité des classes d'affectation C_i générées par la méthode de classement est basée sur sa comparaison avec les classes a priori U_k

n_{ki} est le nombre d'exemples classées dans la classe U_k et ayant été affectés au groupe C_i obtenu par la méthode de classification.
$n_{k.}$ est le nombre d'exemples mises dans la classe a priori U_k
$n_{.i}$ est le nombre d'exemples de la classe C_i
n est le nombre d'exemples.

F mesure

La F-measure combine les mesures de précision et de rappel entre deux classes U_k et C_i de deux partitions.
La mesure de rappel est définie par $R(i,k)=n_{ki}/n_{k.}$
C'est le pourcentage d'exemples de la classe a priori *k* que l'on retrouve dans la classe *i* obtenue par classification.
Le rappel diminue quand le nombre de classes de la partition obtenue par classification diminue.
La mesure de précision est définie par $P(i,k)= n_{ki}/n_{.i}$
C'est le pourcentage d'exemples de la classe *i* que l'on retrouve dans la classe a priori *k.*
La précision augmente quand le nombre de classes de la partition obtenue par classification diminue.

F-mesure

La F-measure proposée par (Van Rijsbergen, 1979) combine les mesures de précision et de rappel entre U_k et C_i.
La mesure de rappel est définie par $R(i,k)=n_{ik}/n_{k.}$
La mesure de précision est définie par $P(i,k)=n_{ik}/n_{.i}$

La F-measure entre la partition a priori U en K groupes et la partition P par la méthode de classification est :

$$F=\sum_{k=1}^{K}(n_{.k}/n)\max_{j}(2.R(k,j).P(k,j)/(R(k,j)+P(k,j)))$$

F mesure pour la classe a priori k :

$$F(k)=\max_{j}(2.R(k,j).P(k,j)/(R(k,j)+P(k,j)))$$

Résultats d'une méthode de classement

=== Run information ===

Scheme: weka.classifiers.trees.J48 -C 0.25 -M 2
Relation: iris
Instances: 150
Attributes: 5
sepallength
sepalwidth
petallength
petalwidth
class
Test mode: split 66% train, remainder test

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

J48 pruned tree
---

petalwidth <= 0.6: Iris-setosa (50.0)
petalwidth > 0.6
| petalwidth <= 1.7
| | petallength <= 4.9: Iris-versicolor (48.0/1.0)
| | petallength > 4.9
| | | petalwidth <= 1.5: Iris-virginica (3.0)
| | | petalwidth > 1.5: Iris-versicolor (3.0/1.0)
| petalwidth > 1.7: Iris-virginica (46.0/1.0)

Number of Leaves : 5

Size of the tree : 9

Critères de qualité

```
=== Evaluation on test split ===
=== Summary ===

Correctly Classified Instances          49               96.0784 %
Incorrectly Classified Instances         2                3.9216 %
Kappa statistic                          0.9408
Mean absolute error                      0.0396
Root mean squared error                  0.1579
Relative absolute error                  8.8979 %
Root relative squared error             33.4091 %
Total Number of Instances               51

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall  F-Measure   Class
  1         0          1         1         1        Iris-setosa
  1         0.063      0.905     1         0.95     Iris-versicolor
  0.882     0          1         0.882     0.938    Iris-virginica
```

```
=== Confusion Matrix ===

  a  b  c   <-- classified as
 15  0  0 |  a = Iris-setosa
  0 19  0 |  b = Iris-versicolor
  0  2 15 |  c = Iris-virginica
```

En ligne les classes d'affectation

En colonne les classes a priori

Critères de qualité

```
=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall  F-Measure   Class
  1         0          1         1         1        Iris-setosa
  1         0.063      0.905     1         0.95     Iris-versicolor
  0.882     0          1         0.882     0.938    Iris-virginica
```

TP rate : taux des « vrais positifs » 15/17

FP rate : taux des « faux positifs » 0/34

Precision : 15/15 Rappel : « recall » 15/17

F-Measure : 2*1*0.882/(1+0.882)= 0.938

	R_V	R_{NV}
V	15	2
NV	0	34

Estimation de la qualité d'une règle de décision

Donner une mesure de qualité à une règle de décision c'est réaliser une estimation du taux ou du coût d'erreur de classement que fournira cette règle sur la population.

Ensemble d'apprentissage

C'est sur cet ensemble qu'une méthode de classement construit la règle de décision.

Ensemble test

C'est sur cet ensemble qu'une méthode de classement est validée

Estimation des taux d'erreur de classement

La probabilité d'erreur de classement ERR sur la population:

$$R(\hat{Y}) = \sum_{k=1}^{K} \pi_k \sum_{l \neq k} \Pr(l / k)$$

Le taux d'erreur de classement sur l'ensemble d'apprentissage : (Taux apparent)

Trop optimiste et avec biais

Le taux d'erreur de classement sur l'ensemble test : (Taux actuel)

Sans biais mais il faut un échantillon important

Techniques de rééchantillonnage (1)

Ensemble de données trop petit (taille *n*)

Validation croisée : (*cross-validation*)

• découper l'échantillon en *k* parties de même effectif

•(*k*-1) parts servent d'ensembles d'apprentissage

• la part restante sert d'ensemble test

Ceci est répété *k* fois et le taux d'erreur de classement est la moyenne des taux d'erreur des ensembles test

Si *k=n* (*leave one out*)

Techniques de rééchantillonnage (2)

Tirage avec remise : bootstrap

On tire au hasard et avec remise *n* exemples qui constituent alors un échantillon

On calcule pour chaque tirage α le taux apparent Err_α et le taux d'erreur apparent sur l'échantillon de base ERR_α

D'où le taux d'erreur bootstrap de *k* dans *l* :

$$Err_B(l/k) = ERR(l/k) + 1/\alpha \sum_\alpha \left| ERR_\alpha(l/k) - Err_\alpha(l/k) \right|$$

Experiment : planifier des expériences

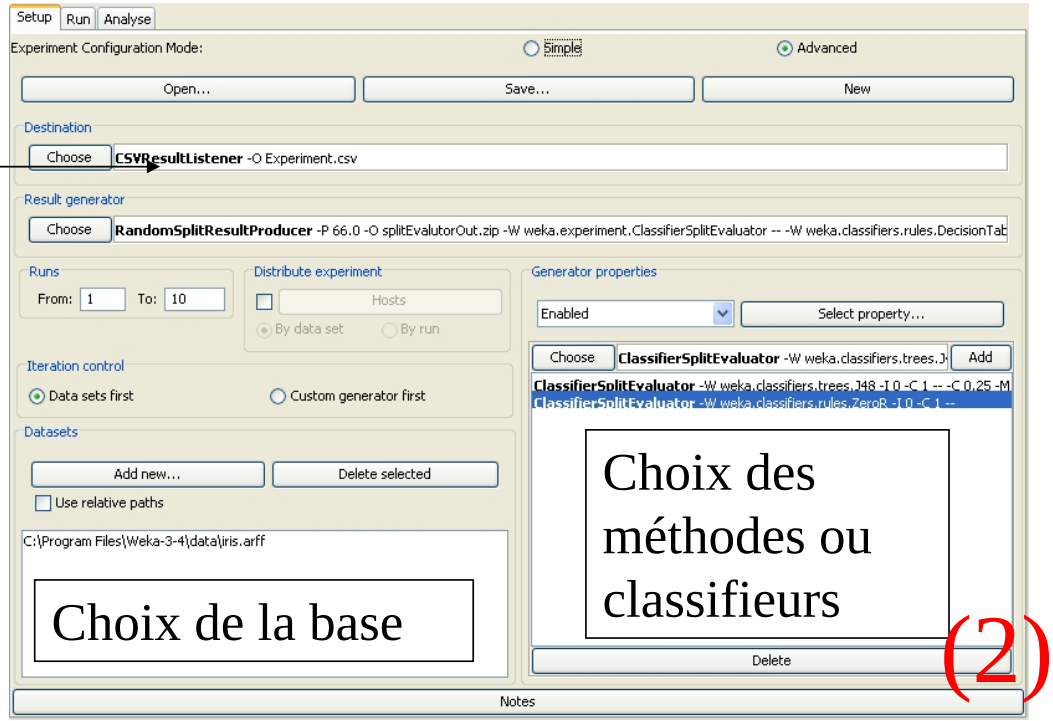
Objectifs :

- Weka Experiment Environment permet créer analyse de comparaison de méthodes (classifiers) ou de stratégies d'utilisation de ces méthodes
- On peut sauvegarder le plan d'expérience et les résultats obtenus
- Une analyse des performances peut être faite via un tableur

Experiment : planifier des expériences

(3)

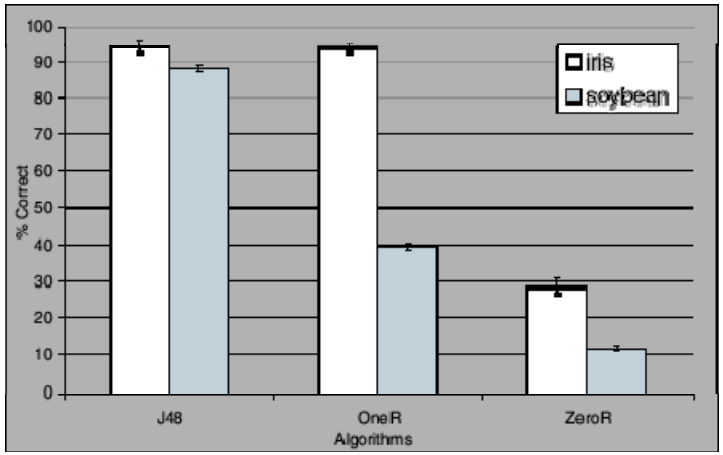
Choix du format de conservation des résultats



(1)

(2)

Experiment : comparer des classifieurs



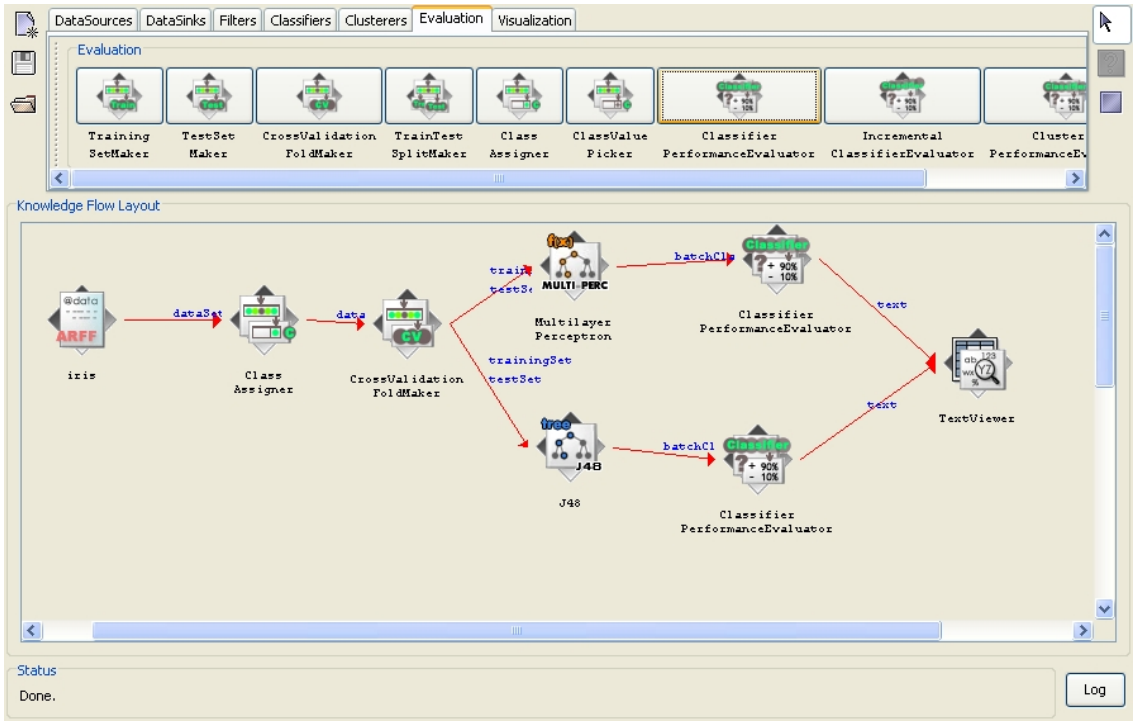
Deux jeux de données et trois méthodes de discrimination

KnowledgeFlow : Créer un enchaînement de modules

Objectif :

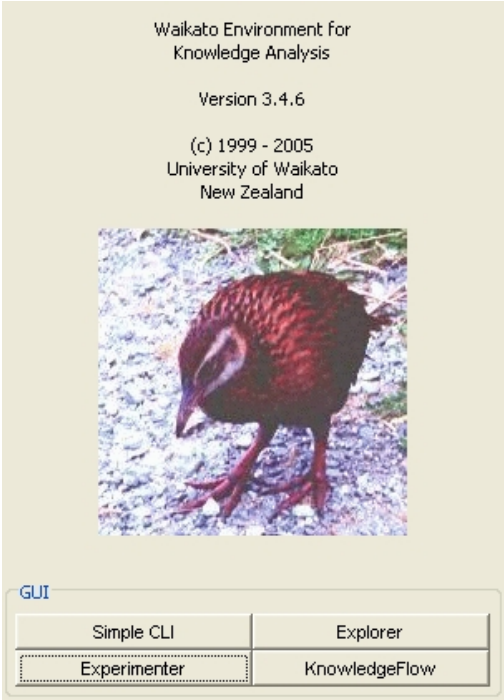
- Créer les liens entre les entrées et sorties de différents modules de manipulation, de visualisation, de décision et d'analyse.
- Permet de créer un traitement complet d'analyse d'un jeu de données
- Programmation « iconique »

KnowledgeFlow : Créer un enchaînement de modules



WEKA:

http://www.cs.waikato.ac.nz/ml/weka/



Copyright: Martin Kramer (mkramer@wxs.nl)