# Algorithmique de l'IA Classification

Ahmed CHADLI Fares GRABA Rémi WATRIGANT Sonia AKROUNE Yasser KADDOUR

20 mai 2010



- Introduction
- 2 Naïve Bayes
- 3 C4.5
- 4 L'implémentation
- 6 Résultats

- Introduction
  - Définitions
  - Critères d'évaluation
- 2 Naïve Bayes
- 3 C4.5
- 4 L'implémentation
- 6 Résultats

## **Définitions**

- $E = E_1 \times ... \times E_n$  ensemble des instances.
- $A \subset E$  ensemble d'apprentissage.
- $T \subset E$  ensemble de test.
- $C = \{c_1, ..., c_k\}$  ensemble des classes.
- $f: E \longrightarrow C$  la fonction d'affectation.

Un classifieur prend en entrée :

- $\{(x, f(x)) : x \in A\}$
- T

Et doit ensuite créer une fonction

$$\hat{f}: T \longrightarrow C$$

## Critères d'évaluation

Une instance  $x \in T$  est bien classée ssi  $\hat{f}(x) = f(x)$ . On mesure alors :

- Pourcentage d'instances de T bien classées (resp. mal classées).
- Pour une classe  $c \in C$ :
  - Faux positifs :  $FP = |\{x \in T : \hat{f}(x) = c \land f(x) \neq c\}|$ . Valeur optimale : 0.
  - Faux négatifs :  $FN = |\{x \in T : \hat{f}(x) \neq c \land f(x) = c\}|$ . Valeur optimale : 0.
  - Vrais positifs :  $TP = |\{x \in T : \hat{f}(x) = f(x) = c\}|$ . Valeur optimale : |T|.
  - Precision :  $\frac{TP}{TP+FP}$ . Valeur optimale : 1.
  - Recall :  $\frac{TP}{TP+FN}$ . Valeur optimale : 1.
  - F-Mesure : 2 \* precision\*recall precision+recall Valeur optimale : 1.

- Introduction
- 2 Naïve Bayes
- 3 C4.5
- 4 L'implémentation
- 6 Résultats

- 1 Introduction
- 2 Naïve Bayes
- 3 C4.5
- 4 L'implémentation
- 6 Résultats

# L'algorithme C4.5

#### En bref:

- Algorithme dû à Ross Quinlan.
- Extension de son précédent algorithme ID3.
- Méthode générant un arbre de décision.

#### Pour construire un noeud de l'arbre :

- Choix d'un attribut qui sépare le mieux l'ensemble d'apprentissage.
- Critère : Entropie relative de chaque attribut.
- L'attribut qui a l'entropie relative la plus forte est choisi pour séparer l'ensemble d'instances.

#### Points forts de C4.5:

- Peut traiter des attributs discrets comme continus.
- Peut traiter des valeurs manquantes.
- Arbres plus petits grâce à un élagage.



- Introduction
- 2 Naïve Bayes
- 3 C4.5
- 4 L'implémentation
- 6 Résultats

# L'implémentation

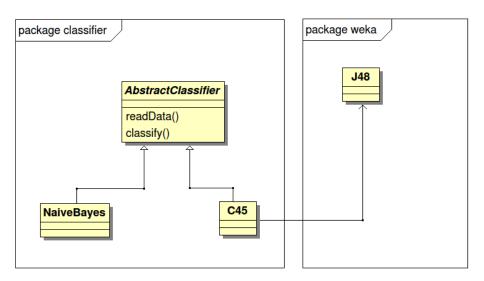
- Langage Java
- Utilise des fichiers .data en entrée
- Utilisation de la librairie Weka pour l'implémentation de C4.5

### Usage :

java -jar classification.jar –algorithm <algorithm> –source <sourceFile> –percentage <percentage> –classIndex <classIndex> –intervalNumber <intervalNumber>

- algorithm: C45 | NaiveBayes: the algorithm to use
- sourceFile : the input data
- percentage : the percentage of data to consider as the training set
- classIndex : the index of the class
- intervalNumber: the number of intervals k





- Introduction
- 2 Naïve Bayes
- 3 C4.5
- 4 L'implémentation
- 6 Résultats

# Résultats

Dataset	num. inst.	nature	num. att.	num. classes	Classification error	
					C4.5	Naive Bayes
balance-scale	625	num	5	3	24.057%	11.321%
cancer	699	num	11	2	5.063%	2.101%
car	1728	mix	7	4	10.562%	13.946%
cmc	1473	num	10	3	46.2%	48.104%
glass	214	num	11	6	1.389%	50.685%
hayes-roth	132	num	6	3	24.444%	46.667%
iris	150	num	5	3	3.922%	5.882%
kr-vs-kp	3196	nom	37	2	0.645%	14.627%
krkopt	28056	mix	7	18	36.125%	64.975%
nursery	12960	mix	9	3	24.013%	25.987%
tic-tac-toe	958	nom	10	2	15.385%	30.982%
transfusion	748	num	5	2	26.772%	27.559%
wine	178	num	14	3	16.667%	29.508%
Z00	101	mix	18	7	14.706%	11.765%
				Mean error	17.853	27.436
num. wins 🐵 11 🖼 🖹 🖼 🔊						