Algorithmique de l'IA
Classification

Ahmed CHADLI
Fares GRABA
Rémi WATRIGANT
Sonia AKROUNE
Yasser KADDOUR

20 mai 2010

Sommaire

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 : $FP = |\{x \in T : \hat{f}(x) = c \wedge f(x) \neq c\}|$.
    Valeur optimale : 0.
  - Faux négatifs : $FN = |\{x \in T : \hat{f}(x) \neq c \wedge f(x) = c\}|$.
    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 : $2 * \frac{precision * recall}{precision + recall}$.
    Valeur optimale : 1.

Sommaire

Théorème de Bayes

A et B deux évènements.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Théorème de Bayes

Indépendance des évènements

$A_1, A_2, ...A_n$ des évènements. Si $A_1, A_2, ...A_n$ sont conditionnellement indépendants alors

$$P(A_1 \cap A_2 \cap ... \cap A_n) = P(A_1).P(A_2)....P(A_n)$$

Théorème de Bayes

Généralisation du théorème de Bayes

Si $A_1, A_2, ...A_n$ sont conditionnellement indépendants alors

$$P(B|A_1 \cap A_2 \cap ... \cap A_n) = \frac{P(A_1|B).P(A_2|B)...P(A_n|B).P(B)}{P(A_1).P(A_2)...P(A_N)}$$

Naïve Bayes

- On calcule dans l'ensemble d'apprentissage A, pour chaque classe $c_i \in C$ la probabilité $P(c = c_i)$
- Pour chaque valeur $e_{j,k}$ de chaque ensemble E_j et pour chaque classe c_i on calcule la probabilité $P(e_j = e_{j,k} | c = c_i)$
- Pour chaque instance $x = (x_1, x_2, ...x_n)$ de T :

$$\hat{f}(x) = \arg\max_{c_i} \{P(c = c_i | e_1 = x_1 \cap e_2 = x_2 \cap ... \cap e_n = x_n)\}$$

avec la formule de Bayes généralisée et en supposant que les e_j sont conditionnellement indépendants.

Sommaire

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.

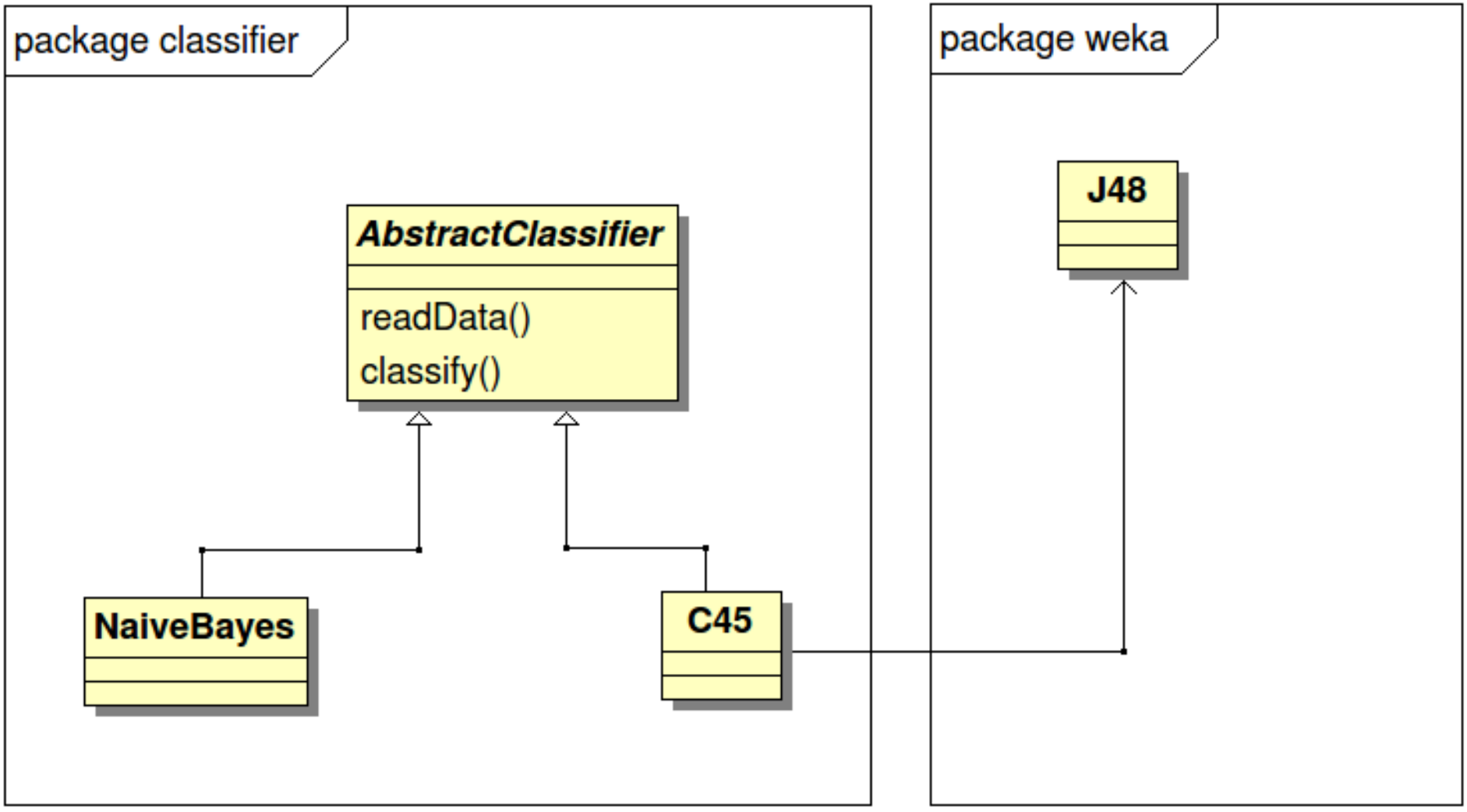
Sommaire

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


package classifier
AbstractClassifier
readData()
classify()
NaiveBayes
C45
package weka
J48

Sommaire

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%
zoo	101	mix	18	7	14.706%	11.765%
				Mean error	17.853	27.436
				num. wins	11	3