

# Contextualisation, visualisation et évaluation en apprentissage non supervisé

Thèse

Laurent  
Candillier

Introduction

État de l'art

Algorithmes

Évaluation  
en cascade

Conclusions

Thèse présentée par  
**Laurent Candillier**

avec le soutien de  
l'équipe GRAppA, université de Lille 3  
et la société Pertinence, Paris

le 15 septembre 2006

# Plan

## Thèse

Laurent  
Candillier

Introduction

État de l'art

Algorithmes

Évaluation  
en cascade

Conclusions

- 1 Introduction
- 2 État de l'art
- 3 Contributions algorithmiques
- 4 Évaluation en cascade
- 5 Conclusions et perspectives

ex : patients atteints de maladies cardio-vasculaires

| id | age | sexe | taille | poids | diabete | cholesterol |
|----|-----|------|--------|-------|---------|-------------|
| 1  | 36  | M    | 180    | 68    | non     | 8           |
| 2  | 16  | M    | 174    | 70    | non     | 16          |
| 3  | 31  | F    | 155    | 49    | oui     | 12          |
| 4  | 11  | F    | 157    | 52    | non     | 11          |
| 5  | 42  | M    | 190    | 82    | oui     | 5           |
| 6  | 35  | F    | 168    | 59    | oui     | 4           |
| 7  | 14  | F    | 149    | 56    | non     | 20          |
| 8  | 18  | M    | 165    | 60    | non     | 15          |
| 9  | 47  | M    | 183    | 79    | oui     | 9           |
| 10 | 35  | F    | 176    | 65    | oui     | 7           |
| 11 | 40  | M    | 170    | 66    | non     | 8           |
| 12 | 10  | F    | 136    | 48    | non     | 22          |
| 13 | 42  | M    | 188    | 98    | non     | 17          |
| 14 | 14  | F    | 146    | 45    | oui     | 16          |

# L'apprentissage non supervisé

Thèse

Laurent  
Candillier

Introduction

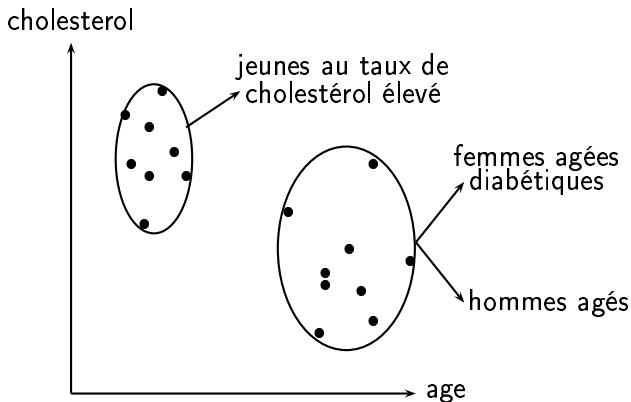
État de l'art

Algorithmes

Évaluation  
en cascade

Conclusions

détection de groupes aux caractéristiques similaires  
*extraction de connaissances*



# L'apprentissage non supervisé

Thèse

Laurent  
Candillier

Introduction

État de l'art

Algorithmes

Évaluation  
en cascade

Conclusions

- 1 définir une distance entre individus  
ex : distance euclidienne dans un espace numérique
- 2 définir une fonction à optimiser
  - minimiser la distance entre individus d'un même groupe
  - maximiser la distance entre individus de groupes distincts
- 3 définir une méthode de recherche de la solution optimale

# Le subspace clustering (*contextualisation*)

Thèse

Laurent  
Candillier

Introduction

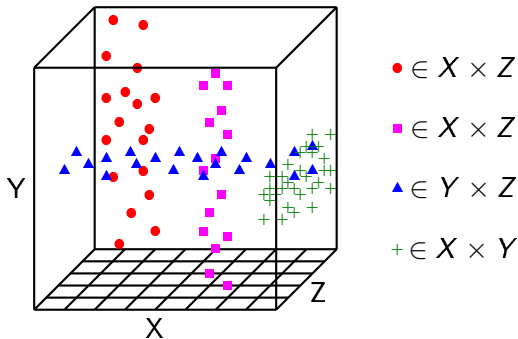
État de l'art

Algorithmes

Évaluation  
en cascade

Conclusions

si tous les attributs n'ont pas la même importance dans la caractérisation des différents groupes, alors une distance définie globalement va échouer



# Comment présenter les résultats ? (*visualisation*)

Thèse

Laurent  
Candillier

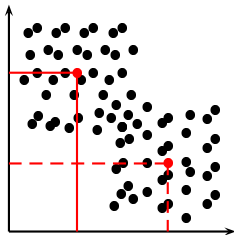
Introduction

État de l'art

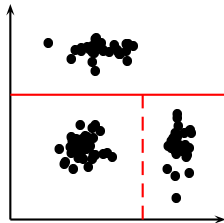
Algorithmes

Évaluation  
en cascade

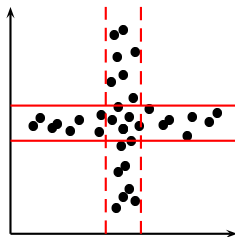
Conclusions



(a) centres



(b) arbre



(c) règles

# Quelle est la solution ? (*évaluation*)

Thèse

Laurent  
Candillier

Introduction

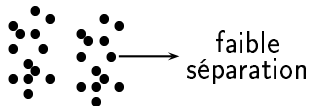
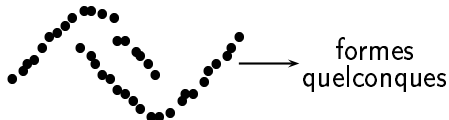
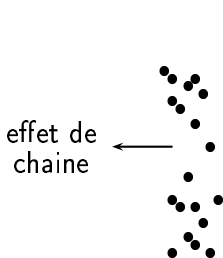
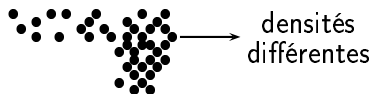
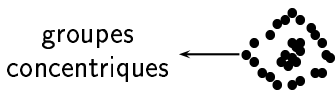
État de l'art

Algorithmes

Évaluation  
en cascade

Conclusions

pour un même jeu de données, différents regroupements  
peuvent être également pertinents



# Plan

Thèse

Laurent  
Candillier

Introduction

État de l'art

Algorithmes

Évaluation  
en cascade

Conclusions

- 1 Introduction
- 2 État de l'art**
- 3 Contributions algorithmiques
- 4 Évaluation en cascade
- 5 Conclusions et perspectives

# K-moyennes [Diday et al., 1982]

Thèse

Laurent  
Candillier

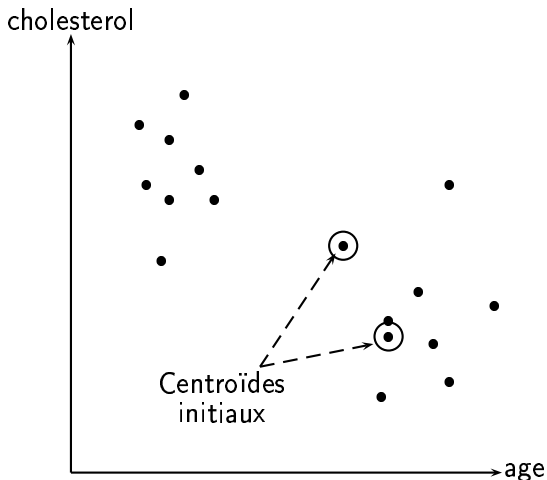
Introduction

État de l'art

Algorithmes

Évaluation  
en cascade

Conclusions



# K-moyennes [Diday et al., 1982]

Thèse

Laurent  
Candillier

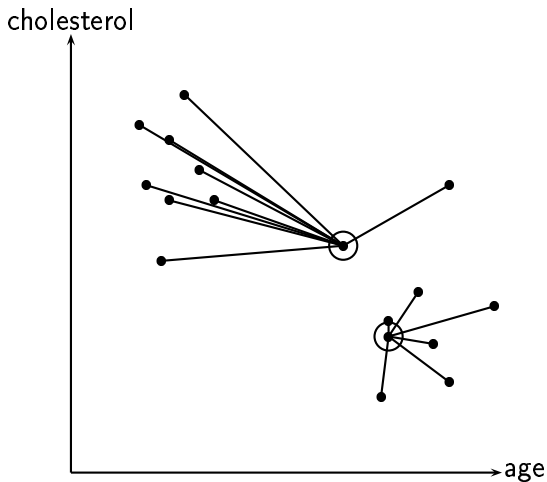
Introduction

État de l'art

Algorithmes

Évaluation  
en cascade

Conclusions



# K-moyennes [Diday et al., 1982]

Thèse

Laurent  
Candillier

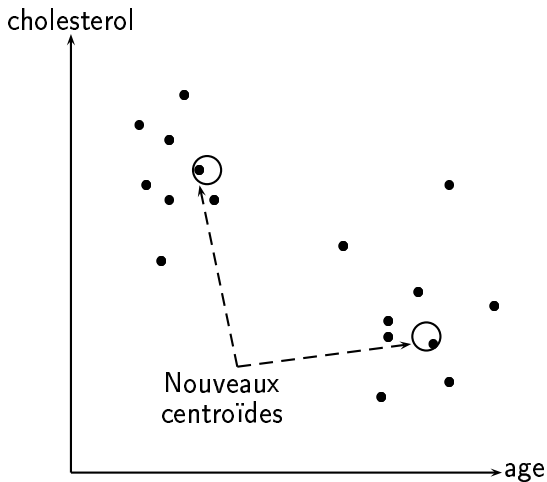
Introduction

État de l'art

Algorithmes

Évaluation  
en cascade

Conclusions



# K-moyennes [Diday et al., 1982]

Thèse

Laurent  
Candillier

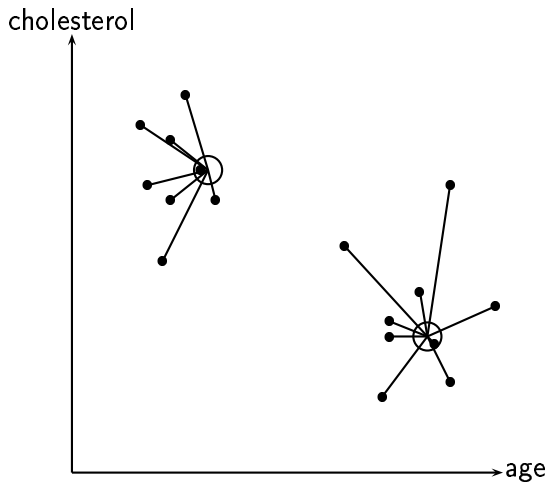
Introduction

État de l'art

Algorithmes

Évaluation  
en cascade

Conclusions



# Méthode paramétrique

Thèse

Laurent  
Candillier

Introduction

État de l'art

Algorithmes

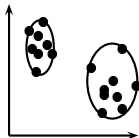
Évaluation  
en cascade

Conclusions

- suppose les données générées par un mélange de distributions connues mais de paramètres inconnus
- objectif : trouver les paramètres de ces distributions en fonction des données
- exemple : distributions gaussiennes représentées par un centre et une matrice de covariance

$$\theta = (\theta_1, \dots, \theta_K) \quad \theta_k = (\vec{\mu}_k, \Sigma_k, \pi_k)$$

- les appartenances des individus aux groupes sont des paramètres cachés du modèle



# Méthode EM [Dempster et al., 1977]

Thèse

Laurent  
Candillier

Introduction

État de l'art

Algorithmes

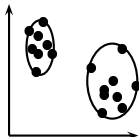
Évaluation  
en cascade

Conclusions

Itérer 2 phases permettant d'améliorer la vraisemblance du modèle par rapport aux données à chaque étape :

- 1 *Estimation* : supposer fixés les paramètres du modèle  $\theta$ , et chercher l'ensemble des affectations des individus aux groupes optimales sous ce modèle
- 2 *Maximisation* : supposer fixées les affectations des individus aux groupes, et calculer les paramètres optimaux du modèle  $\theta$  en fonction de cette connaissance

⇒ généralise K-moyennes



Thèse

Laurent  
Candillier

Introduction

État de l'art

Algorithmes

Évaluation  
en cascade

Conclusions

Méthode ascendante sur les attributs :

- 1 projeter les données sur chaque attribut
- 2 apprentissage non supervisé sur chaque projection
- 3 conserver les attributs contenant des groupes denses

itérer en considérant des projections sur les couples d'attributs conservés, et ainsi de suite jusqu'à ce que plus aucune projection ne contienne de groupe dense

# Subspace clustering descendant

Thèse

Laurent  
Candillier

Introduction

État de l'art

Algorithmes

Évaluation  
en cascade

Conclusions

## PROCLUS [Aggarwal et al., 1999]

- méthode de type K-moyennes
- nécessite le nombre d'attributs pertinents des groupes
- sélectionne pour chaque groupe les attributs sur lesquels les individus sont le moins dispersés

## LAC [Domeniconi et al., 2004]

- méthode de type K-moyennes
- associe à chaque attribut de chaque groupe un poids inversement proportionnel à la dispersion des individus du groupe sur l'attribut

## 1 données artificielles

- comparer à une solution connue a priori
- permet d'observer l'intérêt de la méthode dans des conditions contrôlées
- permet de comparer les résultats de différentes méthodes

## 2 données réelles

- nécessite un expert capable de spécifier si les résultats ont du sens
- permet d'observer l'intérêt de la méthode dans des conditions réelles

## Méthodes de subspace clustering :

- nécessitent des paramètres utilisateur difficiles à spécifier
- ne considèrent que des données numériques
- sensibles à la présence de bruit dans les données
- pas de résultat interprétable

## Méthodes d'évaluation :

- données artificielles : évaluation sur les distributions correspondantes, pas de généralisation aux données réelles
- données réelles : nécessite un expert, pas de généralisation à d'autres données

# Plan

Thèse

Laurent  
Candillier

Introduction

État de l'art

**Algorithmes**

Évaluation  
en cascade

Conclusions

- 1 Introduction
- 2 État de l'art
- 3 Contributions algorithmiques**
- 4 Évaluation en cascade
- 5 Conclusions et perspectives

# Tuareg [Candillier et al., 2004]

Thèse

Laurent  
Candillier

Introduction

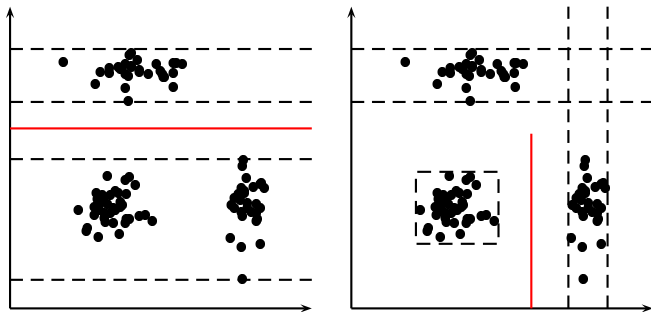
État de l'art

Algorithmes

Évaluation  
en cascade

Conclusions

divisions successives sur les attributs  
à la manière de C4.5 en apprentissage supervisé

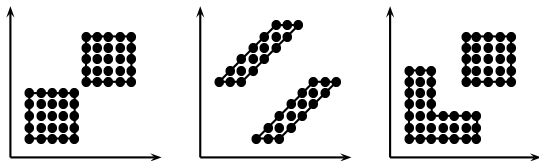


## Atouts :

- ne nécessite aucun paramètre utilisateur
- sélectionne les attributs les plus pertinents
- fournit un résultat interprétable

## Limite :

- ne considère que des données numériques
- n'identifie pas les groupes si leurs différences ne sont visibles sur aucune projection des individus sur 1 attribut



## Principe général :

- adapter les méthodes paramétriques en ajoutant de la sélection d'attributs
- hypothèse d'indépendance des attributs permet de :
  - sélectionner les attributs pertinents
  - prendre en compte différents types d'attributs
  - fournir un résultat compréhensible sous forme de règles
  - accélérer la méthode

## Modèle :

- gaussien sur attribut numérique :  $(\mu_{kd}, \sigma_{kd})$
- multinomial sur attribut catégoriel :  $\overrightarrow{Freqs_{kd}}$
- 2 paramètres :
  - K : nombre de groupes
  - R : nombre d'attributs pertinents des groupes

# Sélection d'attributs

Thèse

Laurent  
Candillier

Introduction

État de l'art

Algorithmes

Évaluation  
en cascade

Conclusions

- 1 associer un poids à chaque attribut de chaque groupe :
  - rapport entre déviation standard locale et globale par rapport au centre du groupe sur les attributs numériques

$$W_{kd} = 1 - \frac{\sigma_{kd}^2}{\Sigma_{kd}^2} \quad \Sigma_{kd}^2 = \frac{1}{N} \sum_i (x_{id} - \mu_{kd})^2$$

- fréquence relative de la catégorie la plus probable du groupe sur les attributs catégoriels

$$W_{kd} = \frac{Freqs_{kd}(cat) - Frequences_d(cat)}{1 - Frequences_d(cat)}$$

$$cat = Argmax_{\{c \in Categories_d\}} Freqs_{kd}(c)$$

- 2 sélectionner les  $R$  attributs de poids maximum pour chaque groupe

NB : sans sélection d'attributs  $\Rightarrow$  SSC [Candillier et al., 2005b]

# Sélection de modèle [Schwartz, 1979]

Thèse

Laurent  
Candillier

Introduction

État de l'art

Algorithmes

Évaluation  
en cascade

Conclusions

Intérêt des méthodes paramétriques :  
spécification des paramètres  $\equiv$  sélection de modèle

- 1 générer plusieurs modèles avec différentes valeurs pour les paramètres  $K$  et  $R$
- 2 sélectionner le modèle le plus approprié aux données  
 $\Rightarrow$  critère statistique BIC :

$$BIC(\theta|D) = -2 \times LL(\theta|D) + M_\theta \times \log N$$

- 1  $LL(\theta|D)$  : vraisemblance du modèle par rapport aux données (critère de la qualité interne du résultat)
- 2  $M_\theta$  : complexité du modèle

compromis entre spécificité et généralisation du modèle

# Présentation des résultats

Thèse

Laurent  
Candillier

Introduction

État de l'art

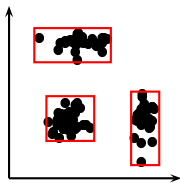
Algorithmes

Évaluation  
en cascade

Conclusions

Sous forme de règles :

- intervalle de définition sur attributs numériques
- catégorie la plus probable sur attributs catégoriels



# Présentation des résultats

Thèse

Laurent  
Candillier

Introduction

État de l'art

Algorithmes

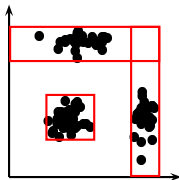
Évaluation  
en cascade

Conclusions

Sous forme de règles :

- intervalle de définition sur attributs numériques
- catégorie la plus probable sur attributs catégoriels

Supprimer les attributs ne modifiant pas le support de la règle



# Présentation des résultats

Thèse

Laurent  
Candillier

Introduction

État de l'art

Algorithmes

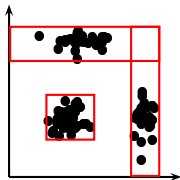
Évaluation  
en cascade

Conclusions

Sous forme de règles :

- intervalle de définition sur attributs numériques
- catégorie la plus probable sur attributs catégoriels

Supprimer les attributs ne modifiant pas le support de la règle



Sélectionner les projections 2D les plus appropriées :

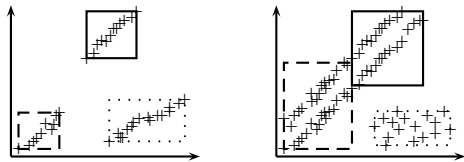
- maximiser la spécificité des règles
- minimiser le recouvrement entre règles

## Atouts :

- ne nécessite aucun paramètre utilisateur
- sélectionne les attributs les plus pertinents
- fournit un résultat interprétable
- capable de considérer différents types d'attributs
- capable de gérer le bruit présent dans les données
- ne souffre pas des limites de Tuareg

## Limite :

- dans certains cas où l'hypothèse d'indépendance des attributs n'est pas respectée



## 1 Méthodes évaluées

- K-moyennes [Diday et al., 1982]
- LAC [Domeniconi et al., 2004]
- SSC [Candillier et al., 2005b]
- SuSE [Candillier et al., 2006b]

## 2 Données artificielles

- mettre en avant l'intérêt du critère BIC pour la sélection de modèle dans SuSE
- observer la robustesse des méthodes face aux attributs non pertinents
- observer la robustesse des méthodes face au bruit présent dans les données

## 3 Données réelles

- données UCI [Blake and Merz, 1998]
- données Pertinence
- données XML [Denoyer et al., 2006]

# Protocole d'expérimentations

Thèse

Laurent  
Candillier

Introduction

État de l'art

Algorithmes

Évaluation  
en cascade

Conclusions

- 1 données artificielles :
  - paramètres : nombre d'individus  $N$ , d'attributs  $M$ , de groupes  $L$ , d'attributs pertinents des groupes  $C$ , et pourcentage de bruit  $p$
  - sélectionner aléatoirement  $L$  *points d'ancrage* dans l'espace de dimension  $M$ , et un nombre proche de  $C$  d'attributs
  - sur les attributs pertinents des groupes, les valeurs des individus sont générées selon des distributions gaussiennes ou multinomiales
  - sur les attributs non pertinents des groupes, les valeurs des individus sont générées selon des distributions uniformes
  - ajout du bruit : individus répartis uniformément dans l'espace de description
  - minimiser l'entropie entre solutions
- 2 données étiquetées : ignorer l'attribut de classe
- 3 données non étiquetées : mener une expertise

# Intérêt du critère BIC pour SuSE

Thèse

Laurent  
Candillier

Introduction

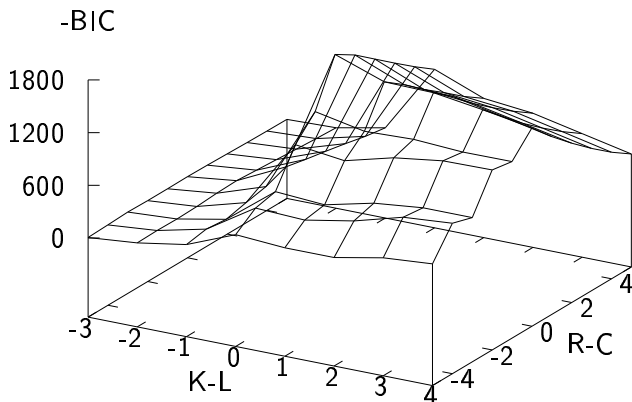
État de l'art

Algorithmes

Évaluation  
en cascade

Conclusions

- 1 K : nombre de groupes / L réel
- 2 R : nombre d'attributs pertinents des groupes / C réel



# Robustesse face aux attributs non pertinents

Thèse

Laurent  
Candillier

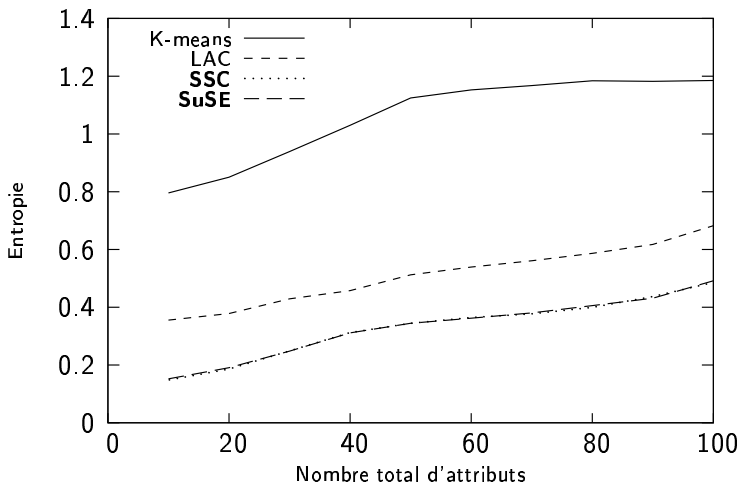
Introduction

État de l'art

Algorithmes

Évaluation  
en cascade

Conclusions



# Temps d'exécution selon le nombre d'attributs

Thèse

Laurent  
Candillier

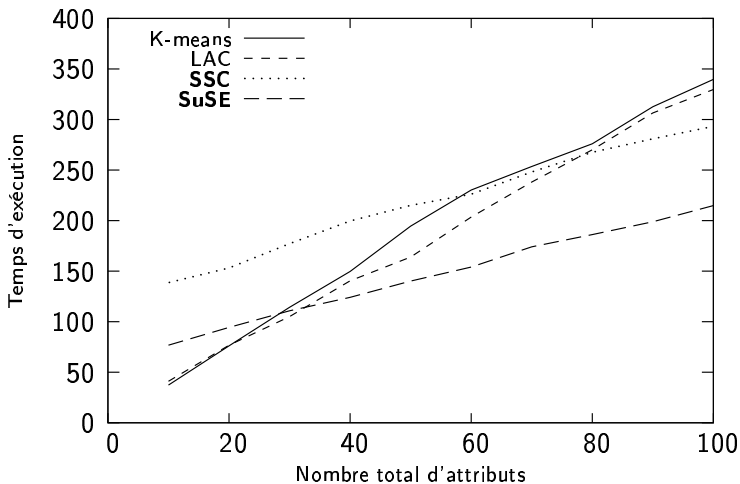
Introduction

État de l'art

Algorithmes

Évaluation  
en cascade

Conclusions



# Robustesse face au bruit

Thèse

Laurent  
Candillier

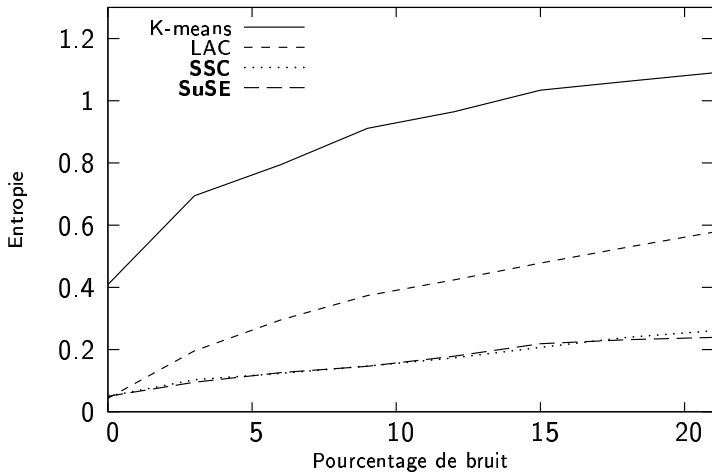
Introduction

État de l'art

Algorithmes

Évaluation  
en cascade

Conclusions



# Données wine [Blake and Merz, 1998]

Thèse

Laurent  
Candillier

Introduction

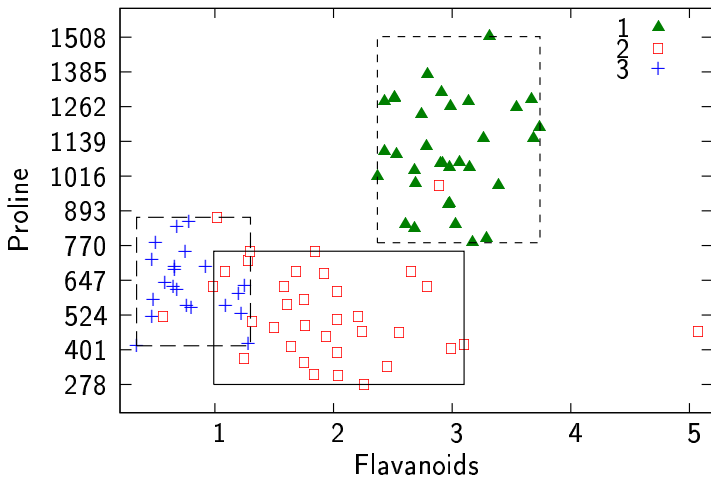
État de l'art

Algorithmes

Évaluation  
en cascade

Conclusions

descriptions de 3 types de vins / 13 attributs



# Données Automobile [Blake and Merz, 1998]

Thèse

Laurent  
Candillier

Introduction

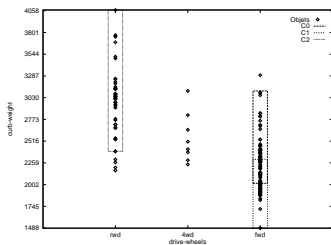
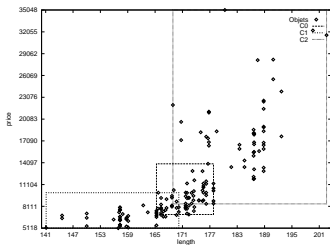
État de l'art

Algorithmes

Évaluation  
en cascade

Conclusions

données non étiquetées / 26 attributs



- prix des voitures augmente fort lorsque leur longueur  $> 170$
- voitures à traction arrière ont un poids à vide supérieur
- majorité des voitures les plus chères sont à traction arrière

# Données Pertinence

Thèse

Laurent  
Candillier

Introduction

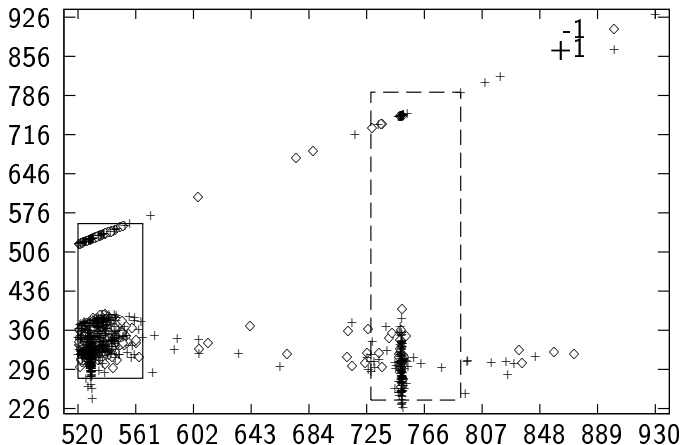
État de l'art

Algorithmes

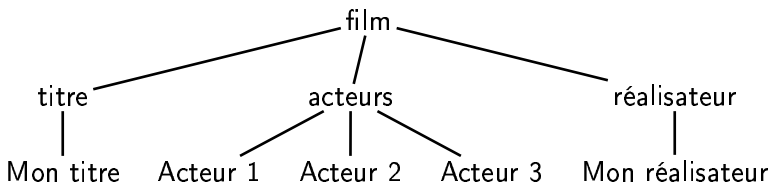
Évaluation  
en cascade

Conclusions

données issues d'un processus industriel / 24 attributs



## Classification de documents XML à partir de leur structure



Proposition : transformation des arbres en attributs-valeurs  
[Candillier et al., 2005a] :

- 1 nombre d'occurrences des labels
- 2 nombre d'occurrences des relations père-fils
- 3 nombre d'occurrences des relations frère-suivant
- 4 nombre d'occurrences des chemins et sous-chemins depuis la racine
- 5 arité des nœuds

# Données XML [Denoyer et al., 2006]

Thèse

Laurent  
Candillier

Introduction

État de l'art

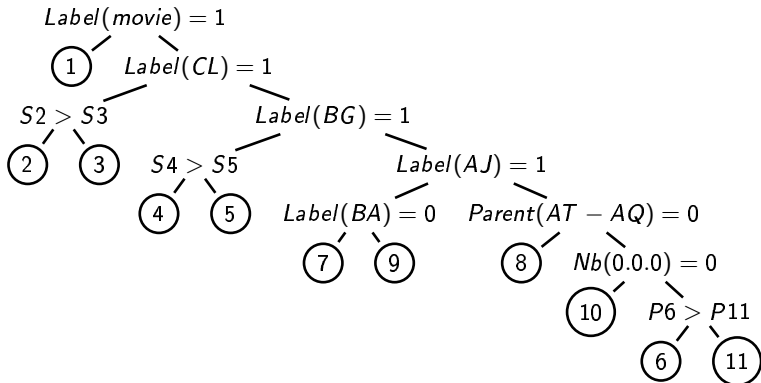
Algorithmes

Évaluation  
en cascade

Conclusions

Comparer les résultats aux classes connues a priori

- 1 adaptation de SuSE en non supervisé : 2/6, compréhensible
- 2 C5 boosté en supervisé : 1/9, robuste au bruit
- 3 adaptation de SuSE en supervisé : 5/9, compréhensible



# Plan

Thèse

Laurent  
Candillier

Introduction

État de l'art

Algorithmes

**Évaluation  
en cascade**

Conclusions

- 1 Introduction
- 2 État de l'art
- 3 Contributions algorithmiques
- 4 Évaluation en cascade**
- 5 Conclusions et perspectives

# Limites des méthodes existantes d'évaluation

Thèse

Laurent  
Candillier

Introduction

État de l'art

Algorithmes

Évaluation  
en cascade

Conclusions

- données artificielles :
  - permet d'observer l'intérêt de la méthode dans des conditions contrôlées
  - permet de comparer les résultats de différentes méthodes
  - mais évaluation sur les distributions correspondantes, pas de généralisation aux données réelles
- données réelles :
  - permet d'observer l'intérêt de la méthode dans des conditions réelles
  - mais nécessite un expert, pas de généralisation à d'autres données

# Évaluation en cascade [Candillier et al., 2006a]

Thèse

Laurent  
Candillier

Introduction

État de l'art

Algorithmes

Évaluation  
en cascade

Conclusions

Proposition : considérer l'apprentissage non supervisé comme un prétraitement à une tâche que l'on sait évaluer

- 1 apprentissage supervisé sur un jeu de données étiqueté
- 2 apprentissage supervisé sur le même jeu de données enrichi par les résultats de l'apprentissage non supervisé
- 3 comparaison des erreurs des 2 classifieurs appris

⇒ si la méthode supervisée est améliorée lorsqu'elle est aidée par les résultats de l'apprentissage non supervisé, alors celui-ci a fourni une information nouvelle et utile

⇒ évaluation plus globale, objective et quantitative

# Méthodes d'enrichissement

Thèse

Laurent  
Candillier

Introduction

État de l'art

Algorithmes

Évaluation  
en cascade

Conclusions

- 1 adapter la *cascade generalization* [Gama and Brazdil, 2000] au cas où un apprenant est non supervisé :
  - apprentissages non supervisés sur le jeu de données pour  $K \in [2, 10]$  en ignorant les informations de classes
  - création de nouveaux attributs pour chaque résultat en fonction des groupes associés aux individus
  - apprentissage supervisé sur le jeu de données enrichi
- 2 apprentissages supervisés indépendants [Apte et al., 2002]
  - apprentissages non supervisés sur le jeu de données pour  $K \in [2, 10]$  en ignorant les informations de classes
  - apprentissages supervisés indépendants sur chaque groupe pour chaque résultat
  - sélection du modèle qui minimise l'erreur en validation croisée

# Mesures de comparaison

Thèse

Laurent  
Candillier

Introduction

État de l'art

Algorithmes

Évaluation  
en cascade

Conclusions

Comparer sur plusieurs jeux de données indépendants les résultats de l'apprenant supervisé avec ou sans l'information issue de l'apprentissage non supervisé

- 1 nb vict : nombre de victoires de l'apprenant supervisé avec ou sans l'information issue de l'apprentissage non supervisé
- 2 vict sign : nombre de victoires significatives selon le  $5 \times 2cv$  *F-test* [Alpaydin, 1999]
- 3 wilcoxon signed rank test : indique si une méthode est significativement meilleure qu'une autre sur un ensemble de problèmes indépendants
- 4 moyenne : erreur pondérée moyenne

- 4 algorithmes supervisés :
  - C4.5 [Quinlan, 1993]
  - C5 boosté [Quinlan, 2004]
  - DLG [Webb and Agar, 1992]
  - SVM [Tsochantaridis et al., 2005]
- 5 algorithmes non supervisés de complexité croissante :  
Aléa, K-moyennes, LAC, SSC, SuSE
- 10 jeux de données issus de l'UCI [Blake and Merz, 1998] :  
sur chaque jeu, 5 validations croisées avec découpage en 2  
[Dietterich, 1998]

# Taux d'erreurs avec C4.5 et enrichissement 1

Thèse

Laurent  
Candillier

Introduction

État de l'art

Algorithmes

Évaluation  
en cascade

Conclusions

|       | C4.5<br>seul | C4.5<br>+ Aléa | C4.5<br>+ K-moyennes | C4.5<br>+ LAC | C4.5<br>+ SSC | C4.5<br>+ SuSE |
|-------|--------------|----------------|----------------------|---------------|---------------|----------------|
| ecoli | 48.5         | 48.3           | 42.8                 | 40.3          | 42            | 43.1           |
| glass | 32.6         | 40.8           | 35.7                 | 37            | 40.4          | 34.9           |
| image | 4.8          | 6              | 4.8                  | 4.6           | 4.6           | 4.6            |
| iono  | 14.1         | 15.8           | 14.2                 | 13.1          | 9.8           | 11.2           |
| iris  | 7.3          | 7.9            | 6.7                  | 3.7           | 5.1           | 4.7            |
| pima  | 31           | 35             | 32.1                 | 32.1          | 30.8          | 30             |
| sonar | 31           | 35.2           | 30                   | 28.8          | 28.8          | 27.2           |
| vowel | 29.5         | 38.5           | 25                   | 26.4          | 24.1          | 22.2           |
| wdbc  | 5.9          | 6.8            | 4.6                  | 3.9           | 5.1           | 3.1            |
| wine  | 8.7          | 8.8            | 10.4                 | 9.6           | 2.7           | 3.6            |

# Taux d'erreurs avec SVM et enrichissement 2

Thèse

Laurent  
Candillier

Introduction

État de l'art

Algorithmes

Évaluation  
en cascade

Conclusions

|       | SVM<br>seul | SVM<br>+ Aléa | SVM<br>+ K-moyennes | SVM<br>+ LAC | SVM<br>+ SSC | SVM<br>+ SuSE |
|-------|-------------|---------------|---------------------|--------------|--------------|---------------|
| ecoli | 52.9        | 63.6          | 36.5                | 35.9         | 40.1         | 36.8          |
| glass | 3.0         | 10.3          | 6.3                 | 5.4          | 4.3          | 4.4           |
| image | 4.9         | 7.8           | 4.7                 | 4.5          | 4.4          | 4             |
| iono  | 20.2        | 23.3          | 15.5                | 18.9         | 13.9         | 14.6          |
| iris  | 4.4         | 4.9           | 2.9                 | 2.0          | 2.7          | 2.5           |
| pima  | 41.5        | 37.3          | 33.1                | 35           | 32           | 31.2          |
| sonar | 28.4        | 30.3          | 22.2                | 20.2         | 18.1         | 17.8          |
| vowel | 55.2        | 59.1          | 24.3                | 24.7         | 25.5         | 26.2          |
| wdbc  | 13.8        | 15.2          | 4                   | 5.2          | 3.8          | 3.7           |
| wine  | 6.8         | 9.3           | 6.8                 | 6.8          | 2.5          | 3.3           |

# Bilan avec C4.5 et enrichissement 1

Thèse

Laurent  
Candillier

Introduction

État de l'art

Algorithmes

Évaluation  
en cascade

Conclusions

|           | C4.5<br>seul | C4.5<br>+ Aléa | C4.5<br>+ K-moyennes | C4.5<br>+ LAC | C4.5<br>+ SSC | C4.5<br>+ SuSE |
|-----------|--------------|----------------|----------------------|---------------|---------------|----------------|
| nb vict   | -            | 1/9            | 5/4                  | 7/3           | 9/1           | 9/1            |
| vict sign | -            | 0/1            | 0/0                  | 1/0           | 2/0           | 3/0            |
| wilcoxon  | -            | -2.67          | -0.05                | 1.31          | 1.83          | 2.56           |
| moyenne   | 21.3         | 24.3           | 20.6                 | 20            | 19.3          | 18.5           |

# Bilan des expérimentations en cascade

Thèse

Laurent  
Candillier

Introduction

État de l'art

Algorithmes

Évaluation  
en cascade

Conclusions

- quels que soient la méthode d'enrichissement et l'algorithme supervisé utilisés, l'ordre dans lequel les méthodes non supervisées sont rangées par l'évaluation en cascade reste toujours le même
- plus le modèle utilisé par l'algorithme non supervisé est complexe, meilleurs sont les résultats
- montre le comportement cohérent de notre méthode
- montre l'intérêt de la sélection d'attributs de SuSE

# Plan

Thèse

Laurent  
Candillier

Introduction

État de l'art

Algorithmes

Évaluation  
en cascade

Conclusions

- 1 Introduction
- 2 État de l'art
- 3 Contributions algorithmiques
- 4 Évaluation en cascade
- 5 Conclusions et perspectives**

Problème : différents groupes peuvent être décrits par différents attributs pertinents (*subspace clustering*)

⇒ intérêt des méthodes paramétriques pour identifier simultanément les groupes et leurs attributs pertinents

+ aide à la présentation compréhensible des résultats

## Cadre : extraction de connaissances

- minimiser le nombre de connaissances a priori requises de la part de l'utilisateur
- fournir un résultat compréhensible

⇒ intérêt des méthodes paramétriques et de l'hypothèse d'indépendance des attributs

⇒ décrire les groupes en utilisant des règles caractérisées par un minimum d'attributs parmi les plus pertinents

+ utiliser la projection en 2 dimensions

Problème : différents regroupements des données peuvent être pertinents pour différentes raisons

⇒ évaluation en cascade :

- comportement cohérent
- évaluation plus globale, objective, quantitative
- met en avant l'intérêt de la sélection des attributs pertinents mise en œuvre dans SuSE

## 1 Contextualisation :

- descente de gradient sur le critère BIC pour SuSE
- validations théoriques
- données symboliques / courbes

## 2 Visualisation :

- en 3 dimensions

## 3 Évaluation :

- tester dans le cadre d'autres tâches que l'on sait évaluer  
ex : vitesse d'accès après une indexation automatique de  
base de données OLAP
- approfondir l'étude dans le cadre de la combinaison de  
classifieurs

# Merci

## Thèse

Laurent  
Candillier

Introduction

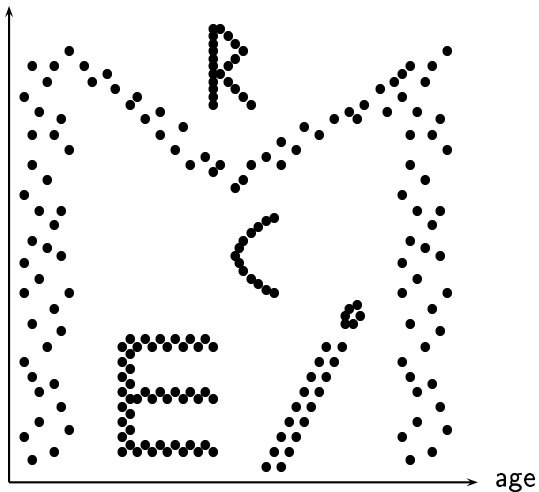
État de l'art




Algorithmes

Évaluation  
en cascade

Conclusions

cholesterol



-  Aggarwal, C. C., Wolf, J. L., Yu, P. S., Procopiuc, C., and Park, J. S. (1999).  
Fast algorithms for projected clustering.  
In ACM SIGMOD International Conference on Management of Data, pages 61–72.
-  Agrawal, R., Gehrke, J., Gunopulos, D., and Raghavan, P. (1998).  
Automatic subspace clustering of high dimensional data for data mining applications.  
In ACM SIGMOD International Conference on Management of Data, pages 94–105, Seattle, Washington.
-  Alpaydin, E. (1999).  
Combined 5x2cv F-test for comparing supervised classification learning algorithms.  
Neural Computation, 11(8) :1885–1892.

 Apte, C. V., Natarajan, R., Pednault, E. P. D., and Tipu, F. A. (2002).

A probabilistic estimation framework for predictive model analytics.

[IBM Systems Journal](#), 41(3).

 Blake, C. and Merz, C. (1998).

UCI repository of machine learning databases

[<http://www.ics.uci.edu/~mlern/MLRepository.html>].

 Candillier, L., Tellier, I., and Torre, F. (2004).

Tuareg : Classification non supervisée contextualisée.

In Liquière, M. and Sebban, M., editors, [6ème Conférence francophone sur l'Apprentissage automatique \(CAp'2004\)](#), pages 159–174.

 Candillier, L., Tellier, I., and Torre, F. (2005a).

Transforming XML trees for efficient classification and clustering.

[INEX 2005 Workshop on Mining XML documents.](#)



Candillier, L., Tellier, I., Torre, F., and Bousquet, O. (2005b).

**SSC : Statistical Subspace Clustering.**





In Perner, P. and Imiya, A., editors, [4th International Conference on Machine Learning and Data Mining in Pattern Recognition \(MLDM'2005\)](#), volume LNAI 3587 of [LNCS](#), pages 100–109, Leipzig, Germany. Springer Verlag.



Candillier, L., Tellier, I., Torre, F., and Bousquet, O. (2006a).

**Évaluation en cascade d'algorithmes de clustering.**

In Miclet, L., editor, [8ème Conférence francophone sur l'Apprentissage automatique \(CAp'2006\)](#), pages 109–124.

-  Candillier, L., Tellier, I., Torre, F., and Bousquet, O. (2006b).  
SuSE : Subspace Selection embedded in an EM algorithm.  
In Miclet, L., editor, 8ème Conférence francophone sur l'Apprentissage automatique (CAp'2006), pages 331–345.
-  Dempster, A., Laird, N., and Rubin, D. (1977).  
Maximum likelihood from incomplete data via the EM algorithm.  
Journal of the Royal Statistical Society, Series B,  
39(1) :1–38.
-  Denoyer, L., Gallinari, P., and Vercoastre, A. M. (2006).  
XML Mining Challenge at INEX 2005.  
Technical report, University of Paris VI, INRIA.
-  Diday, E., Lemaire, J., Pouget, J., and Testu, F. (1982).

Un algorithme de type nuées dynamiques, pages 117–123.  
Dunod.



Dietterich, T. G. (1998).

Approximate statistical test for comparing supervised classification learning algorithms.

Neural Computation, 10(7) :1895–1923.



Domeniconi, C., Papadopoulos, D., Gunopulos, D., and Ma, S. (2004).

Subspace clustering of high dimensional data.

In SIAM International Conference on Data Mining.



Gama, J. and Brazdil, P. (2000).

Cascade generalization.

Machine Learning, 41(3) :315–343.



Quinlan, J. R. (1993).

## C4.5 : Programs for Machine Learning. KAUFM.



Quinlan, J. R. (2004).

Data mining tools see5 and c5.0.



Schwartz, G. (1979).

Estimating the dimension of a model.

[The Annals of Statistics](#), 6(2) :461–464.



Tsochantaridis, I., Joachims, T., Hofmann, T., and Altun, Y. (2005).

Large margin methods for structured and interdependent output variables.

[Journal of Machine Learning Research \(JMLR\)](#), 6 :1453–1484.



Webb, G. I. and Agar, J. W. M. (1992).

Thèse

Laurent  
Candillier

Introduction

État de l'art

Algorithmes

Évaluation  
en cascade

**Conclusions**

Inducing diagnostic rules for glomerular disease with the DLG machine learning algorithm.

[Artificial Intelligence in Medicine](#), 4 :419–430.