Chapitre 7

Validation de la qualité des connaissances

Rappel sur les étapes de l'ECD :

1. Compréhension du domaine d'application
2. Création du fichier cible (target data set)
3. Traitement des données brutes (data cleaning and preprocessing)
4. Réduction des données (data reduction and projection)
5. Définition des tâches de fouille de données
6. Choix des algorithmes appropriés de fouille de données
7. Fouille de données (data mining)
8. Interprétation des formes extraites (mined patterns)
9. Validation des connaissances extraites

Etape de Validation :

Dans cette étape, on distingue 2 modes de validation **:**

- Par expertise (qualitative)
- Statistique (Quantitative)

Pour certains domaines d'application (le diagnostic médical, par exemple), il est essentiel que le modèle produit soit validé selon les 2 modes :

- Première validation du modèle produit par l'expert
- Seconde validation statistique sur des bases de cas existantes

Qualité d'un modèle de DM :

- La qualité d'un modèle de fouille de données obtenu se mesure selon les critères suivants :
  - Il est rapide à créer
  - Il est rapide à utiliser
  - Il est compréhensible pour l'utilisateur
  - Ses performances sont bonnes et ne se dégrade pas dans le temps
  - il est fiable
  - Il évolue facilement.
- Aucun modèle n'a toutes ces qualités
- Il n'existe pas de meilleure méthode de fouille :
  - Il faut faire des compromis selon les besoins dégagés et les caractéristiques connues des outils
  - Pour une utilisation optimale, une combinaison de méthodes est recommandée.

Évaluer et valider les résultats

- Évaluation qualitative (subjective par des experts)
  - Restitution de la connaissance sous forme graphique ou sous une forme interprétable
- Évaluation quantitative (objective par des méthodes statistiques)
  - Notion d'intervalle de confiance (indicateurs pour la pertinence des règles, seuil de confiance et intervalle de confiance fonction de la taille de l'échantillon)
  - Validation par le test (base de test) matrice de confusion / éclairage métier

Évaluer et valider les résultats

- Évaluation quantitative: Matrice de confusion
  - comparaison des cas observés par rapport aux prédictions
    - exemple : prédiction de factures impayées

Prédit ↓	**Observé**			
	Payé	Retardé	Impayé	Total
Payé	80	15	5	100
Retardé	1	17	2	20
Impayé	5	2	23	30
Total	86	34	30	150

  - Validité du modèle
    - nombre exacte (diagonale) / nombre totale = 120/150 = 0.80

Évaluer et valider les résultats

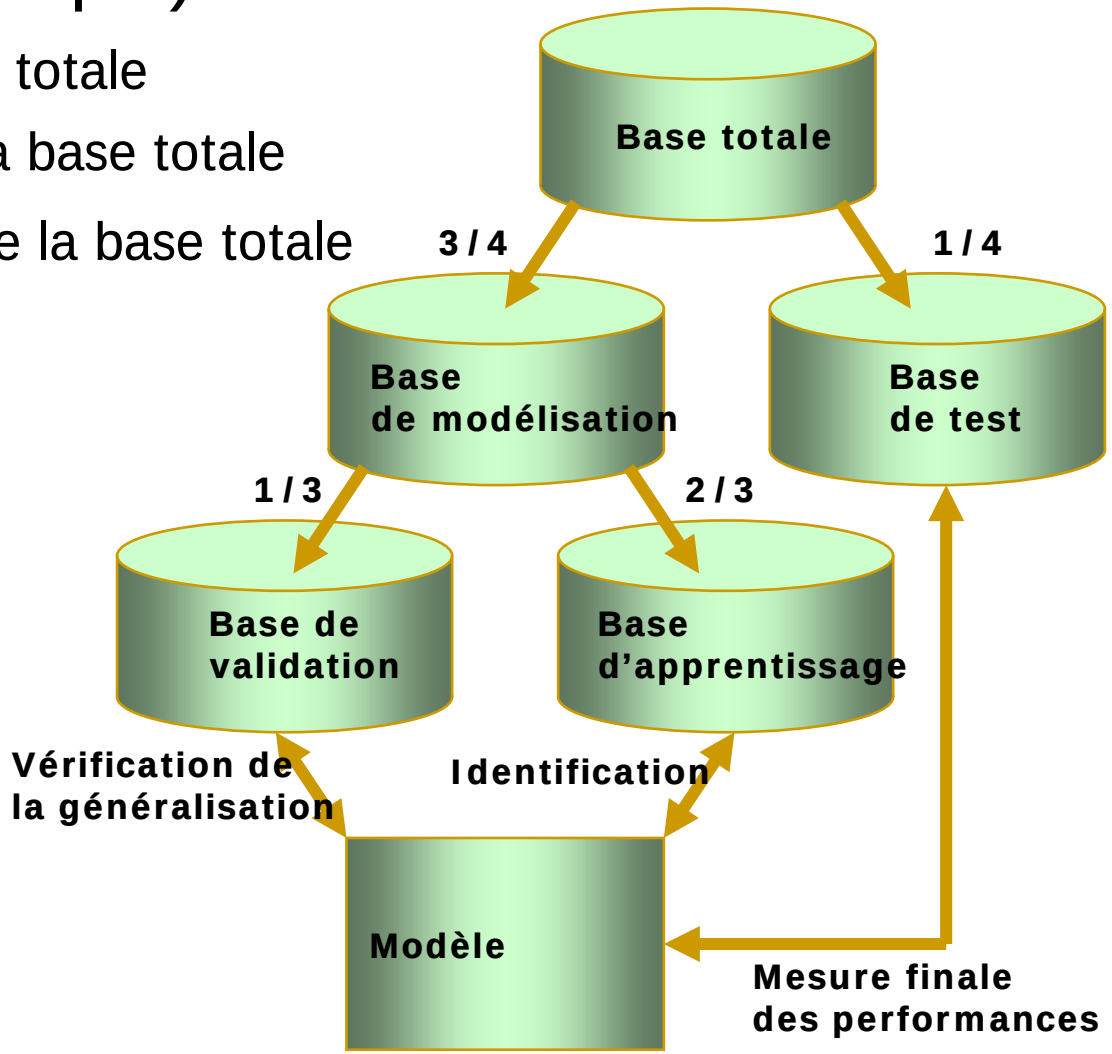
Évaluation par le test:

Pour permettre l'apprentissage, le test et la validation l'échantillon global peut être partagé en trois sous-échantillons qui sont :

- Le « training data set » dont l'objectif est de permettre l'apprentissage de l'algorithme
- Le « test data set » dont l'objectif consiste à éviter le sur ou le sous apprentissage
- Le « validation data set » dont l'objectif consiste à comparer les performances de plusieurs modèles

Évaluer et valider les résultats

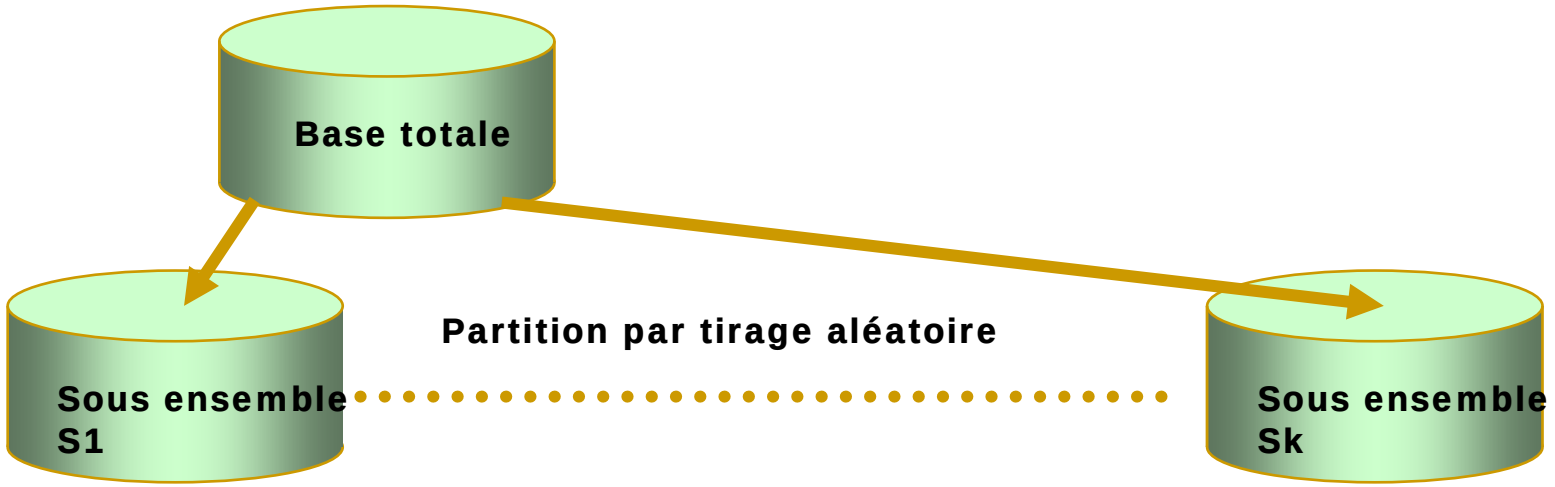
- Évaluation par le test (Exemple)
  - Base de test 25% de la base totale
  - Base de validation 25% de la base totale
  - Base d'apprentissage 50% de la base totale



Évaluer et valider les résultats

- Validation croisée

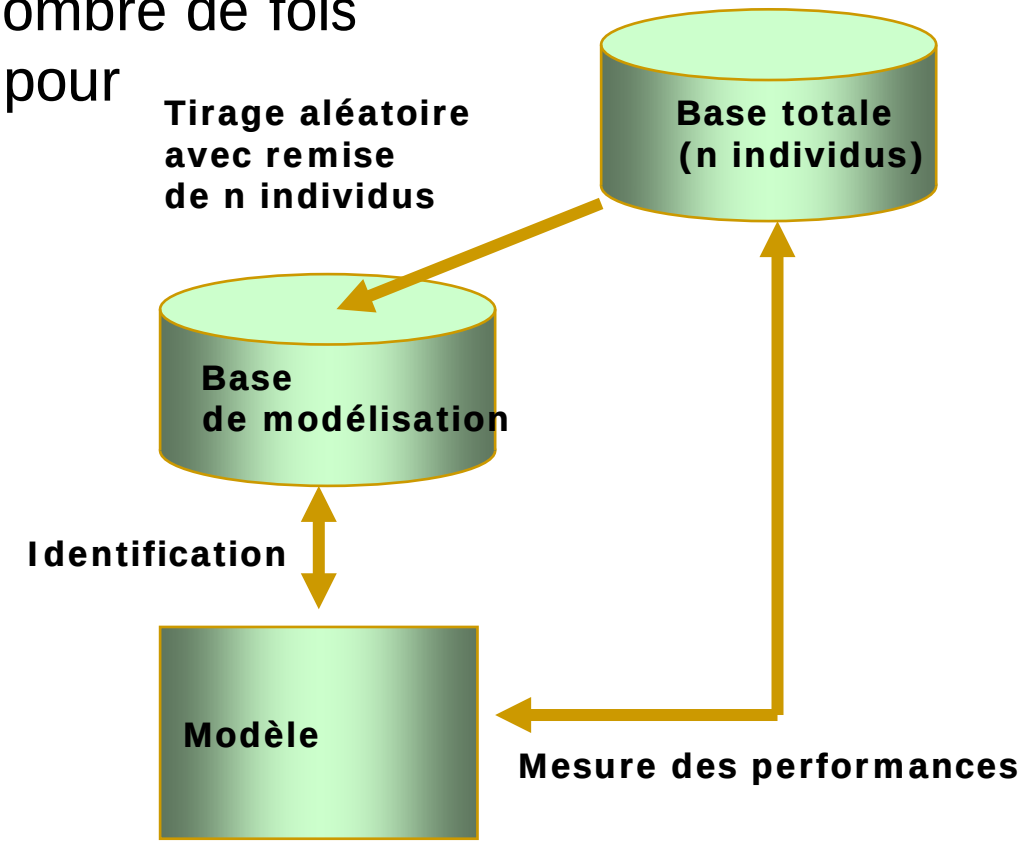
• Diviser les données en k sous ensemble
• Utiliser k-1 sous ensemble comme données d'apprentissage et un sous ensemble comme données de test



Évaluer et valider les résultats

- Bootstrap

n instances testes aléatoires (ensemble de données réduit):
L'expérience est réalisée un grand nombre de fois
et les performances sont agrégées pour
fournir une évaluation globale



Validation et type de méthode

Les méthodes de validation vont déprendre de la nature de la tâche et du problème considéré. Nous distinguerons deux modes de validation : statistique et par expertise.

Pour certains domaines d'application (le diagnostic médical, par exemple), il est essentiel que le modèle produit soit compréhensible. Il y a donc une première validation du modèle produit par l'expert, celle-ci peut être complétée par une validation statistique sur des bases de cas existantes.

Validation des méthodes non supervisées

Pour les problèmes d'apprentissage non supervisé (segmentation, association), la validation est essentiellement du ressort de l'expert. Pour la segmentation, le programme construit des groupes homogènes, un expert peut juger de la pertinence des groupes constitués. La encore, on peut combiner avec une validation statistique sur un problème précis utilisant cette segmentation. Pour la recherche des règles d'association, c'est l'expert du domaine qui jugera de la pertinence des règles, en effet, l'algorithme, s'il fournit des règles porteuses d'information, produit également des règles triviales et sans intérêt.

Validation des méthodes supervisées

Pour la validation statistique, la première tâche à réaliser consiste à utiliser les méthodes de base de statistique descriptive. L'objectif est d'obtenir des informations qui permettront de juger le résultat obtenu, ou d'estimer la qualité ou les biais des données d'apprentissage.
Calculer les moyennes et variances des attributs.
Si possible, calculer la corrélation entre certains champs.
Déterminer la classe majoritaire dans le cas de la classification.

Validation des méthodes supervisées

Pour la classification supervisée, la deuxième tâche consiste à décomposer les données en plusieurs ensembles disjoints. L'objectif est de garder des données pour estimer les erreurs des modèles ou de les comparer. Il est souvent recommandé de constituer:

- Un ensemble d'apprentissage.
- Un ensemble de test.
- Un ensemble de validation

Validation des méthodes supervisées

Au moins deux ensembles sont nécessaires : l'ensemble d'apprentissage permet de générer le modèle, l'ensemble test permet d'évaluer l'erreur réelle du modèle sur un ensemble indépendant évitant ainsi un biais d'apprentissage. Lorsqu'il s'agit de tester plusieurs modèles et de les comparer, on peut sélectionner le meilleur modèle selon ses performances sur l'ensemble test et ensuite évaluer son erreur réelle sur l'ensemble de validation.

Lorsqu'on ne dispose pas de suffisamment d'exemples on peut se permettre d'apprendre et d'estimer les erreurs avec un même ensemble par la technique de *validation croisée.*

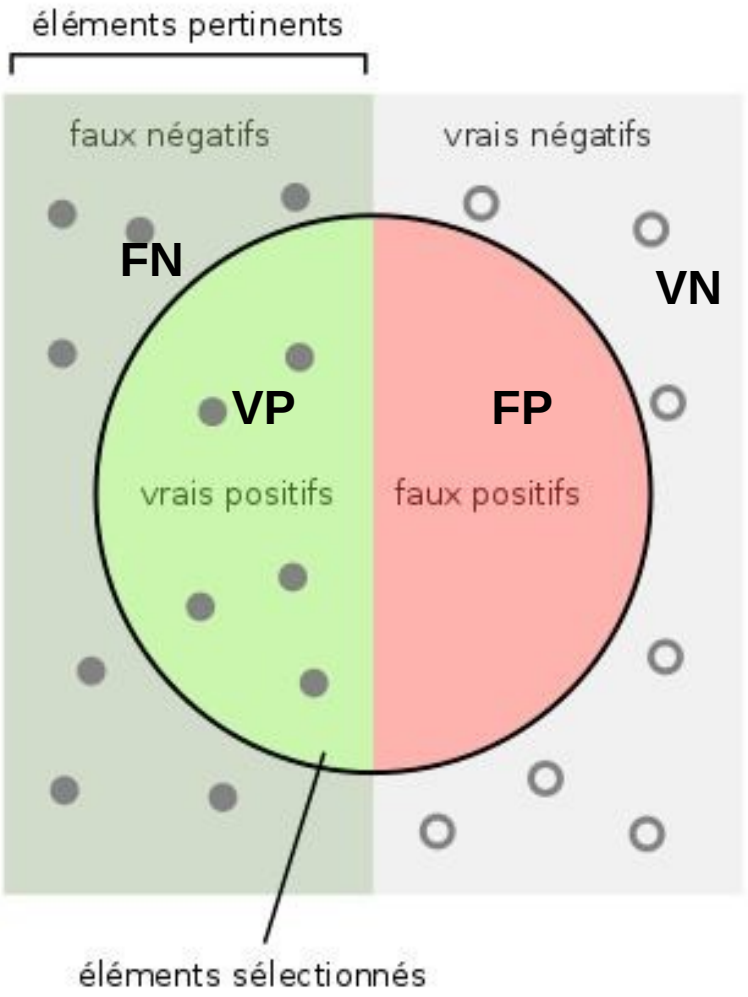
Les métriques d'évaluation

L'apprentissage supervisé utilise une partie des données pour calculer un modèle de décision qui sera généralisé sur l'ensemble du reste de l'espace. Il est très important d'avoir des mesures permettant de qualifier le comportement du modèle appris sur les données non utilisées lors de l'apprentissage. Les métriques sont: Rappel, précision

$$\text{Précision} = \frac{\text{Le nombre de données correctement attribués à la classe i}}{\text{Le nombre total des données attribués à la classe i}}$$

$$\text{Rappel} = \frac{\text{Le nombre de données correctement attribués à la classe i}}{\text{Le nombre total de données appartenant à la classe i}}$$

Les métriques d'évaluation



Combien de candidats sélectionnés sont pertinents ?

$$\text{Précision} = \frac{\text{vrais positifs}}{\text{éléments sélectionnés}}$$

Combien d'éléments pertinents sont sélectionnés ?

$$\text{Rappel} = \frac{\text{vrais positifs}}{\text{éléments pertinents}}$$

Matrice de confusion

Une matrice de confusion montre le nombre de prédictions correctes et incorrectes faites par le modèle de classification par rapport à la valeur réelle (valeur cible) dans les données. La matrice est NxN, où N est le nombre de valeurs cibles (classes). Les performances de ces modèles sont généralement évaluées à l'aide des données de la matrice. Le tableau suivant affiche une matrice de confusion 2x2 pour deux classes (positive et négative).

Matrice de confusion

Matrice de confusion		Cible			
		Positive	Négative		
Modéle	Positive	a (VP)	b (VN)	Valeur prédictive positive	a/(a+b)
	Négative	c (FP)	d (FN)	Valeur prédictive négative	d/(c+d)
		Sensibilité	Spécificité	**Précision** = (a+d)/(a+b+c+d)	
		a/(a+c)	d/(b+d)		

Précision : la proportion du nombre total de prédictions qui étaient correctes.
Valeur Prédictive Positive ou Précision : la proportion de cas positifs qui ont été correctement identifiés.
Valeur Prédictive Négative : la proportion de cas négatifs correctement identifiés.
Sensibilité ou Rappel : la proportion de cas positifs réels qui sont correctement identifiés.
Spécificité : la proportion de cas réels négatifs correctement identifiés.

Exemple

Nous disposons des deux bases d'exemples d'entrainement et de test suivantes:

Base d'entrainement

N°	Age	Niv	Trav	Vote
1	V	U	T	Oui
2	J	U	T	Oui
3	V	M	C	Non
4	J	M	C	Non
5	J	P	T	Oui
6	J	P	C	Oui
7	V	P	C	Non
8	V	U	C	Non
9	J	U	C	Oui
10	J	M	C	Oui

Base de test

N°	Age	Niv	Trav	Vote
1	J	U	C	Oui
2	V	P	T	Non
3	V	M	T	Oui
4	J	M	T	Non
5	J	P	C	Oui

On souhaite construire des modèles de décision et les tester afin de les utiliser pour la prédiction du prochain vote.

Exemple

1. Méthode ZéroR

ZeroR est le classificateur le plus simple, il se base sur la classe et ignore tous les attribut. Il prédit tout simplement la classe majoritaire. Il n'a aucune capacité de prédiction mais représente une base de comparaison pour les autres méthodes.

- Construire un modèle de décision en utilisant la méthode ZéroR.
- Donner sa table de confusion, précision et rappel sur la base de test

Exemple

2. **Méthode OneR**

OneR (One Rule) est un algorithme simple et précis de classification. Il génère une règle pour chaque attribut dans les données puis sélectionne la règle d'erreur minimale comme la seule règle de décision.

- Construire un modèle de décision en utilisant la méthode OneR.
- Calculer les erreurs.

Exemple

Algorithme OneR:

Algorithme OneR

pour chaque attribut **faire**

pour chaque valeur de cet attribut **faire** Créer une règle comme suit :

Calculer le nombre d'apparition de chaque valeur de la classe ;

Trouver la classe la plus fréquente ;

Créer une règle assignant cette classe à cette valeur de l'attribut

fin pour

fin pour

Calculer l'erreur totale de chaque règle pour chaque attribut

Choisir l'attribut de l'erreur minimale

Exemple

Classification par analyse des règles d'association

Dans la classification associative (par règles d'association), les règles d'association sont générées et analysées pour les utiliser en classification. L'idée est de rechercher les règles solides contenant dans leur partie droite l'attribut classe, c-à-d de la forme :

$$\text{Attribut1} = v_{att1} \wedge \text{Attribut2} = v_{att2} \wedge \ldots \wedge \text{Attributn} = v_{attn} \Rightarrow \text{Classe} = v_{classe}$$

Plusieurs études ont montré que cette technique est plus précise que certaines méthodes traditionnelles tel que les arbres de décision.

Exemple

3. Algorithme CBA

CBA (Classification Based Association)

L'algorithme CBA est l'un des premiers algorithmes de classification associative. Il utilise l'algorithme Apriori pour générer les règles d'association puis utilise une heuristique pour construire le classificateur. Les règles sont ordonnées selon leurs supports et confidences. Si plusieurs règles ont la même partie gauche, la règle de la confidence la plus élevée est utilisée dans le classificateur. Pour classer un nouveau tuplet, la première règle le satisfaisant est utilisée. Le classificateur contient aussi une règle par défaut pour classer les tuplet dont une règles satisfaisante n'existe pas.

Exemple

3. Algorithme CBA

Les étapes de l'algorithme CBA sont:

- Trouver les MFs (Motifs ou itemsets fréquents)
- Trouver les règles solides
- Construire le modèle
- Construire la table de confusion
- Calculer la précision et le rappel