Introduction à la fouille de données

M. Ledmi
m_ledmi@esi.dz

Département d'Informatique Khenchela

2020/2021

Plan

Vous êtes ici

Motivation : Le besoin crée l'invention

Problème de l'explosion de données !

- Les outils automatiques de collecte de données font que les bases de données contiennent énormément de données.
- Plusieurs sources de données :
  - Entrepôts du Web : ex. Google, youtube.
  - Réseaux sociaux et hébergement de documents : ex. Facebook, gmail.
  - e-commerce : Achats dans les supermarchés, transactions de cartes bancaires

Motivation : Le besoin crée l'invention

Problème de l'explosion de données !

- Les outils automatiques de collecte de données font que les bases de données contiennent énormément de données.
- Plusieurs sources de données :
  - Entrepôts du Web : ex. Google, youtube.
  - Réseaux sociaux et hébergement de documents : ex. Facebook, gmail.
  -

Motivation : Le besoin crée l'invention

Problème de l'explosion de données !

- Les outils automatiques de collecte de données font que les bases de données contiennent énormément de données.
- Plusieurs sources de données :
  - Entrepôts du Web : ex. Google, youtube.
  - Réseaux sociaux et hébergement de documents : ex. Facebook, gmail.
  - e-commerce : Achats dans les supermarchés, transactions de cartes bancaires

Motivation : Le besoin crée l'invention

Problème de l'explosion de données !

- Les outils automatiques de collecte de données font que les bases de données contiennent énormément de données.
- Plusieurs sources de données :
  - Entrepôts du Web : ex. Google, youtube.
  - Réseaux sociaux et hébergement de documents : ex. Facebook, gmail.
  - e-commerce : Achats dans les supermarchés, transactions de cartes bancaires

Motivation : Le besoin crée l'invention

Problème de l'explosion de données !

- Les outils automatiques de collecte de données font que les bases de données contiennent énormément de données.
- Plusieurs sources de données :
  - Entrepôts du Web : ex. Google, youtube.
  - Réseaux sociaux et hébergement de documents : ex. Facebook, gmail.
  - e-commerce : Achats dans les supermarchés, transactions de cartes bancaires

Motivation : Le besoin crée l'invention

Problème de l'explosion de données !

- Les outils automatiques de collecte de données font que les bases de données contiennent énormément de données.
- Plusieurs sources de données :
  - Entrepôts du Web : ex. Google, youtube.
  - Réseaux sociaux et hébergement de documents : ex. Facebook, gmail.
  - e-commerce : Achats dans les supermarchés, transactions de cartes bancaires

Beaucoup de données mais peu de connaissances !

- Difficulté d'accès à l'information.
- Trop de pistes à explorer.

Motivation : Le besoin crée l'invention

Solution !

Par analogie à la *recherche des pépites d'or* dans un gisement, la **fouille de données** vise à :

- Extraire des informations cachées par analyse globale ;
- Découvrir des modèles ("patterns") difficiles à percevoir car :
  - Le volume de données est très grand
  - Le nombre de variables à considérer est important
  - Ces "patterns" sont imprévisibles (même à titre d'hypothèse à vérifier).

Motivation : Le besoin crée l'invention

Solution !

Par analogie à la *recherche des pépites d'or* dans un gisement, la **fouille de données** vise à :

- Extraire des informations cachées par analyse globale ;
- Découvrir des modèles ("patterns") difficiles à percevoir car :
  - Le volume de données est très grand
  - Le nombre de variables à considérer est important

Motivation : Le besoin crée l'invention

Solution !

Par analogie à la *recherche des pépites d'or* dans un gisement, la **fouille de données** vise à :

- Extraire des informations cachées par analyse globale ;
- Découvrir des modèles ("patterns") difficiles à percevoir car :
  - Le volume de données est très grand
  - Le nombre de variables à considérer est important
  - Ces "patterns" sont imprévisibles (même à titre d'hypothèse à vérifier).

Motivation : Le besoin crée l'invention

Solution !

Par analogie à la *recherche des pépites d'or* dans un gisement, la **fouille de données** vise à :

- Extraire des informations cachées par analyse globale ;
- Découvrir des modèles ("patterns") difficiles à percevoir car :
  - Le volume de données est très grand
  - Le nombre de variables à considérer est important
  - Ces "patterns" sont imprévisibles (même à titre d'hypothèse à vérifier).

Motivation : Le besoin crée l'invention

Solution !

Par analogie à la *recherche des pépites d'or* dans un gisement, la **fouille de données** vise à :

- Extraire des informations cachées par analyse globale ;
- Découvrir des modèles ("patterns") difficiles à percevoir car :
  - Le volume de données est très grand
  - Le nombre de variables à considérer est important
  - Ces "patterns" sont imprévisibles (même à titre d'hypothèse à vérifier).

Evolution des Bases de Données

Historique

- **1960s** : Collecte des données, création des BD's.
- 1970s : Modèle et SGBD's relationnels, SQL, transactions.
- **1980s** : Modèles de données et SGBD's avancés (relationnel étendu, OO, déductifs, etc.) et SGBD's dédiés (spatial, génomique, engineering, etc.)

Evolution des Bases de Données

Historique

- **1960s** : Collecte des données, création des BD's.
- 1970s : Modèle et SGBD's relationnels, SQL, transactions.
- **1980s** : Modèles de données et SGBD's avancés (relationnel étendu, OO, déductifs, etc.) et SGBD's dédiés (spatial, génomique, engineering, etc.)
- **1990s** : Data mining et data warehousing, BD's multimédia, BD's sur le WEB

Evolution des Bases de Données

Historique

- **1960s** : Collecte des données, création des BD's.
- 1970s : Modèle et SGBD's relationnels, SQL, transactions.
- **1980s** : Modèles de données et SGBD's avancés (relationnel étendu, OO, déductifs, etc.) et SGBD's dédiés (spatial, génomique, engineering, etc.)
- **1990s** : Data mining et data warehousing, BD's multimédia, BD's sur le WEB

Evolution des Bases de Données

Historique

- **1960s** : Collecte des données, création des BD's.
- **1970s** : Modèle et SGBD's relationnels, SQL, transactions.
- **1980s** : Modèles de données et SGBD's avancés (relationnel étendu, OO, déductifs, etc.) et SGBD's dédiés (spatial, génomique, engineering, etc.)
- **1990s** : Data mining et data warehousing, BD's multimédia, BD's sur le WEB

Concept-clé : Donnée

Donnée

Une donnée est le résultat direct d'une mesure.

- Elle peut être collectée par un outil de supervision, par une personne ou être déjà présente dans une base de données par ex.
- Une donnée seule ne permet pas de prendre une décision sur une action à lancer.

Concept-clé : Donnée

Donnée

Une donnée est le résultat direct d'une mesure.

- Elle peut être collectée par un outil de supervision, par une personne ou être déjà présente dans une base de données par ex.
- Une donnée seule ne permet pas de prendre une décision sur une action à lancer.

Concept-clé : Donnée

Donnée

Une donnée est le résultat direct d'une mesure.

- Elle peut être collectée par un outil de supervision, par une personne ou être déjà présente dans une base de données par ex.
- Une donnée seule ne permet pas de prendre une décision sur une action à lancer.

Exemple :

- Il fait 15° dans cette pièce.

Concept-clé : Information

Information

Une information est une donnée à laquelle un sens et une interprétation ont été donnés.

- Une information permet à un responsable opérationnel de prendre une décision (d'échelle locale ou à petite échelle) sur une action à mener.

Concept-clé : Information

Information

Une information est une donnée à laquelle un sens et une interprétation ont été donnés.

- Une information permet à un responsable opérationnel de prendre une décision (d'échelle locale ou à petite échelle) sur une action à mener.

Concept-clé : Information

Information

Une information est une donnée à laquelle un sens et une interprétation ont été donnés.

- Une information permet à un responsable opérationnel de prendre une décision (d'échelle locale ou à petite échelle) sur une action à mener.

Exemple :

les données précédentes sont interprétées de la manière suivante :

- Il fait froid dans cette pièce.

Concept-clé : Connaissance

Connaissance

La connaissance est le résultat d'une réflexion sur les informations analysées en se basant sur :

- ses expériences, ses idées, ses valeurs.
- les avis d'autres personnes consultées pour l'occasion

Concept-clé : Connaissance

Connaissance

La connaissance est le résultat d'une réflexion sur les informations analysées en se basant sur :

- ses expériences, ses idées, ses valeurs.
- les avis d'autres personnes consultées pour l'occasion

Concept-clé : Connaissance

Connaissance

La connaissance est le résultat d'une réflexion sur les informations analysées en se basant sur :

- ses expériences, ses idées, ses valeurs.
- les avis d'autres personnes consultées pour l'occasion

Exemple :

- Pour avoir chaud, il suffit de monter le chauffage.

Quelques références bibliographiques

- **Data Mining : Concepts and techniques**,
  - *Auteur* :Jiawei Han & Micheline Kamber,
  - *Edition* : Morgan Kaufmann, 2000.
- **Fouille de données, Notes de cours**,
  - *Auteur* : Ph. PREUX, Université de Lille 3
  - *Lien* : `http:\www.grappa.univ-lille3.fr/~ppreux/fouille`

Vous êtes ici

Introduction

La révolution numérique a rendu l'information facile à être

- capturer, traiter, stocker,
- distribuer et transmettre.

Progrès et utilisation des technologies informatiques dans les différents domaines de la vie,

- Grandes quantités de données diverses continuent d'être collectées et stockées dans les bases de données,
- Si la quantité d'informations double tous les mois, la taille et le nombre de bases de données augmente proportionnellement à un rythme similaire.

Introduction

La révolution numérique a rendu l'information facile à être

- capturer, traiter, stocker,
- distribuer et transmettre.

Progrès et utilisation des technologies informatiques dans les différents domaines de la vie,

- Grandes quantités de données diverses continueront d'être collectées et stockées dans les bases de données.
- Si la quantité d'informations double tous les mois, la taille et le nombre de bases de données augmente probablement à un rythme similaire.

Introduction

La révolution numérique a rendu l'information facile à être

- capturer, traiter, stocker,
- distribuer et transmettre.

Progrès et utilisation des technologies informatiques dans les différents domaines de la vie,

- Grandes quantités de données diverses continueront d'être collectées et stockées dans les bases de données.
- Si la quantité d'informations double tous les mois, la taille et le nombre de bases de données augmente probablement à un rythme similaire.

Introduction

La révolution numérique a rendu l'information facile à être

- capturer, traiter, stocker,
- distribuer et transmettre.

Progrès et utilisation des technologies informatiques dans les différents domaines de la vie,

- Grandes quantités de données diverses continueront d'être collectées et stockées dans les bases de données.
- Si la quantité d'informations double tous les mois, la taille et le nombre de bases de données augmente probablement à un rythme similaire.

Introduction

La révolution numérique a rendu l'information facile à être

- capturer, traiter, stocker,
- distribuer et transmettre.

Progrès et utilisation des technologies informatiques dans les différents domaines de la vie,

- Grandes quantités de données diverses continueront d'être collectées et stockées dans les bases de données.
- Si la quantité d'informations double tous les mois, la taille et le nombre de bases de données augmente probablement à un rythme similaire.

Introduction

L'extraction des connaissances à partir de ce grand volume est un défi :

- Plus on a de données,
- Plus il est difficile d'en tirer de la connaissance.

La fouille de données est une tentative

- Explorer et d'analyser ces énormes volume de données afin d'y découvrir de l'information implicite.
- Règles d'association, une classification ou une segmentation de population.

Introduction

L'extraction des connaissances à partir de ce grand volume est un défi :

- Plus on a de données,
- Plus il est difficile d'en tirer de la connaissance.

La fouille de données est une tentative

- Explorer et d'analyser cet énorme volume de données afin d'y découvrir de l'information implicite.
- Règles d'association, une classification ou une segmentation de population.

Introduction

L'extraction des connaissances à partir de ce grand volume est un défi :

- Plus on a de données,
- Plus il est difficile d'en tirer de la connaissance.

La fouille de données est une tentative

- Explorer et d'analyser cet énorme volume de données afin d'y découvrir de l'information implicite.
- Règles d'association, une classification ou une segmentation de population.

Introduction

L'extraction des connaissances à partir de ce grand volume est un défi :

- Plus on a de données,
- Plus il est difficile d'en tirer de la connaissance.

La fouille de données est une tentative

- Explorer et d'analyser cet énorme volume de données afin d'y découvrir de l'information implicite.
- Règles d'association, une classification ou une segmentation de population.

Introduction

L'extraction des connaissances à partir de ce grand volume est un défi :

- Plus on a de données,
- Plus il est difficile d'en tirer de la connaissance.

La fouille de données est une tentative

- Explorer et d'analyser cet énorme volume de données afin d'y découvrir de l'information implicite.
- Règles d'association, une classification ou une segmentation de population.

Fouille de données

Data mining

L'extraction des connaissances à partir des données est un processus non trivial d'identification des modèles valides, nouveaux, potentiellement utiles et au final compréhensibles, à partir de données.

Fouille de données

Data mining

L'extraction des connaissances à partir des données est un processus non trivial d'identification des modèles valides, nouveaux, potentiellement utiles et au final compréhensibles, à partir de données.

- **Valide** : vérifiée par des experts du domaine et correcte dans le futur.
- **Nouveau** : Ce qui est recherché est non prévisible, inconnu.
- **Utile** : utilisé pour prendre des décisions.

Fouille de données

Data mining

L'extraction des connaissances à partir des données est un processus non trivial d'identification des modèles valides, nouveaux, potentiellement utiles et au final compréhensibles, à partir de données.

- **Valide** : vérifiée par des experts du domaine et correcte dans le futur.
- **Nouveau** : Ce qui est recherché est non prévisible, inconnu.
- **Utile** : utilisé pour prendre des décisions.
- **Compréhensible** : significatif et facile à comprendre.

Fouille de données

Data mining

L'extraction des connaissances à partir des données est un processus non trivial d'identification des modèles valides, nouveaux, potentiellement utiles et au final compréhensibles, à partir de données.

- **Valide** : vérifiée par des experts du domaine et correcte dans le futur.
- **Nouveau** : Ce qui est recherché est non prévisible, inconnu.
- **Utile** : utilisé pour prendre des décisions.
- **Compréhensible** : significatif et facile à comprendre.

Fouille de données

Data mining

L'extraction des connaissances à partir des données est un processus non trivial d'identification des modèles valides, nouveaux, potentiellement utiles et au final compréhensibles, à partir de données.

- **Valide** : vérifiée par des experts du domaine et correcte dans le futur.
- **Nouveau** : Ce qui est recherché est non prévisible, inconnu.
- **Utile** : utilisé pour prendre des décisions.
- **Compréhensible** : significatif et facile à comprendre.

Schéma de l'ECD

1. La compréhension du domaine d'application :

- Connaissance a priori, de l'application.
- Connaissance des objectifs à atteindre.

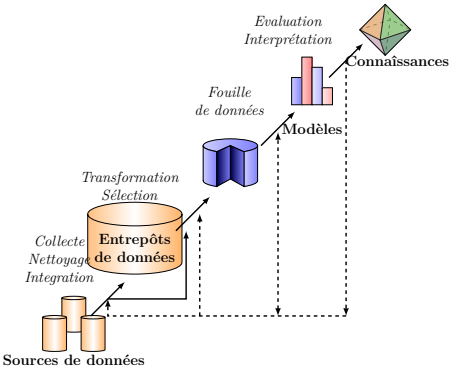


Schéma de l'ECD

1. La compréhension du domaine d'application :

- Connaissance a priori, de l'application.
- Connaissance des objectifs à atteindre.

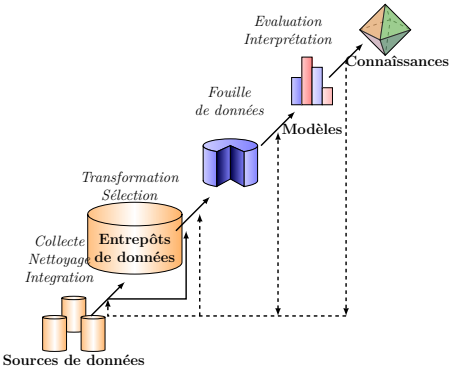


Schéma de l'ECD

1. La compréhension du domaine d'application :

- Connaissance a priori, de l'application.
- Connaissance des objectifs à atteindre.

2. Extractions des données cibles :

- Sélection d'un ensemble de données.
- Concentrer sur un sous-ensemble de variables.

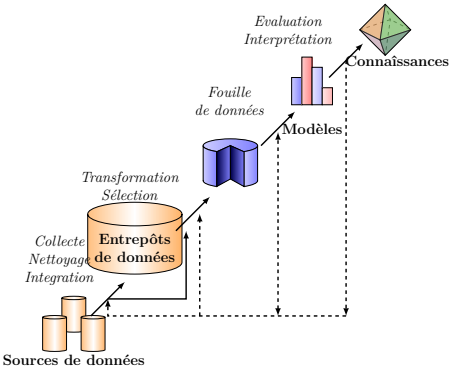


Schéma de l'ECD

1. La compréhension du domaine d'application :

- Connaissance a priori, de l'application.
- Connaissance des objectifs à atteindre.

2. Extractions des données cibles :

- Sélection d'un ensemble de données.
- Concentrer sur un sous-ensemble de variables.

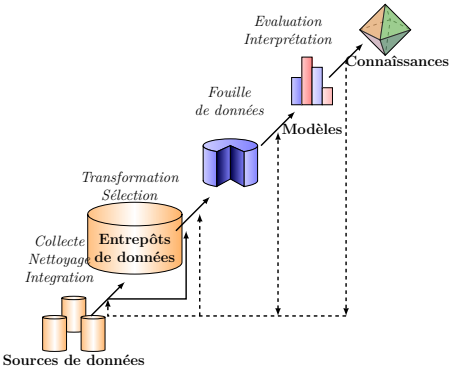


Schéma de l'ECD

3. Prétraitements des données :

- améliore la qualité des données.
- augmente l'efficacité de l'extraction.
- *Nettoyage* : normalisation, suppression du bruit , manipulation des données manquantes.
- *Intégration* : multiples ensembles de données hétérogènes.

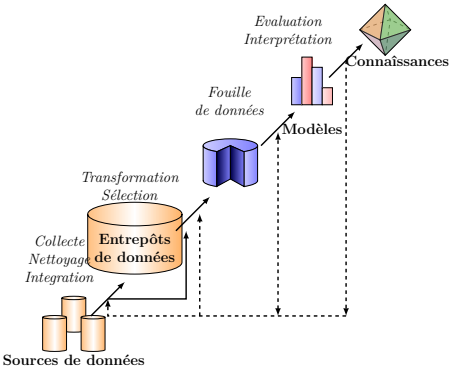


Schéma de l'ECD

3. Prétraitements des données :

- améliore la qualité des données.
- augmente l'efficacité de l'extraction.
- *Nettoyage* : normalisation, suppression du bruit , manipulation des données manquantes.
- *Intégration* : multiples ensembles de données hétérogènes.

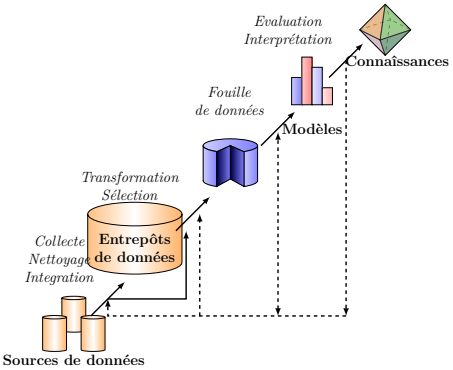


Schéma de l'ECD

3. Prétraitements des données :

- améliore la qualité des données.
- augmente l'efficacité de l'extraction.
- *Nettoyage* : normalisation, suppression du bruit , manipulation des données manquantes.
- *Intégration* : multiples ensembles de données hétérogènes.

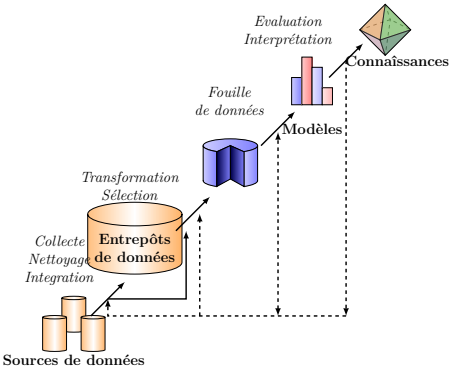


Schéma de l'ECD

3. Prétraitements des données :

- améliore la qualité des données.
- augmente l'efficacité de l'extraction.
- *Nettoyage* : normalisation, suppression du bruit , manipulation des données manquantes.
- *Intégration* : multiples ensembles de données hétérogènes.

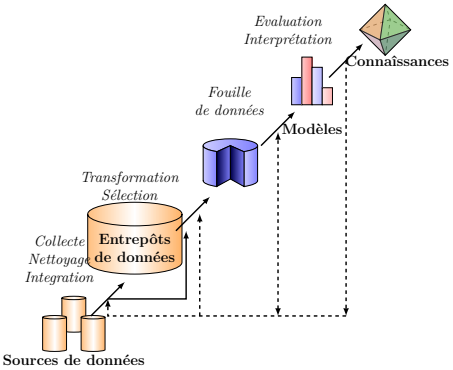


Schéma de l'ECD

4. Fouille de données : correspond à l'une ou plusieurs des tâches :

- Classification,
- Clustering,
- Règles d'association , . . . etc.

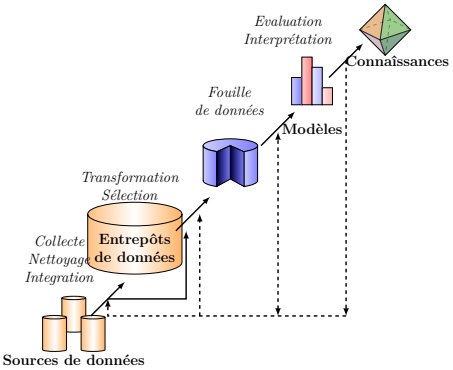


Schéma de l'ECD

4. Fouille de données : correspond à l'une ou plusieurs des tâches :

- Classification,
- Clustering,
- Règles d'association , . . . etc.

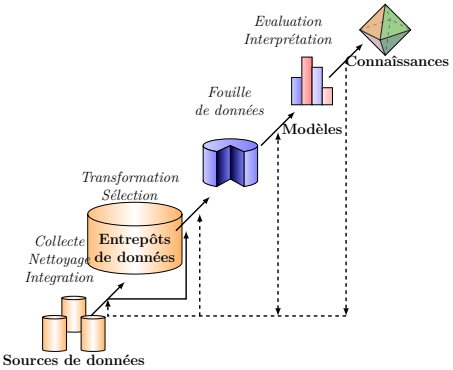


Schéma de l'ECD

4. Fouille de données : correspond à l'une ou plusieurs des tâches :

- Classification,
- Clustering,
- Règles d'association , . . . etc.

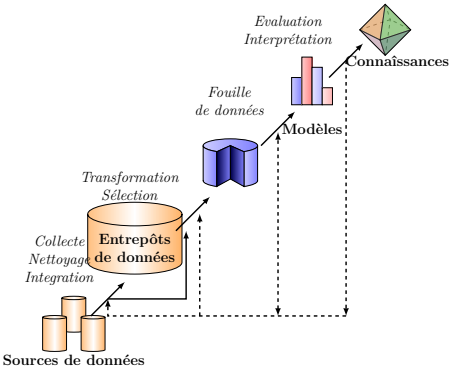


Schéma de l'ECD

5. Interprétation :

- Interprétation des modèles découverts,
- Visualisation possible des modèles extraits.
- Evaluer les modèles extraits pour identifier les modèles utiles pour l'utilisateur.

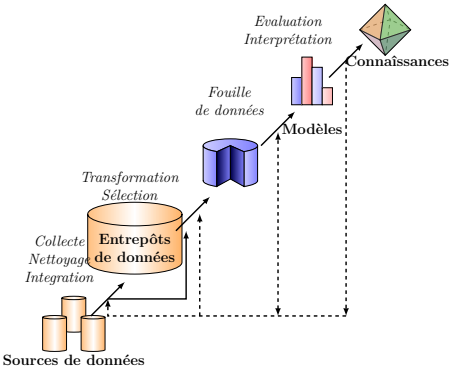


Schéma de l'ECD

5. Interprétation :

- Interprétation des modèles découverts,
- Visualisation possible des modèles extraits.
- Evaluer les modèles extraits pour identifier les modèles utiles pour l'utilisateur.

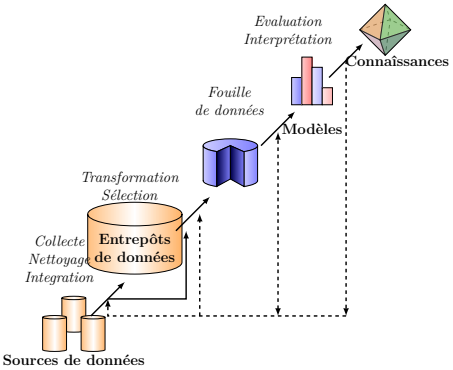


Schéma de l'ECD

5. Interprétation :

- Interprétation des modèles découverts,
- Visualisation possible des modèles extraits.
- Evaluer les modèles extraits pour identifier les modèles utiles pour l'utilisateur.

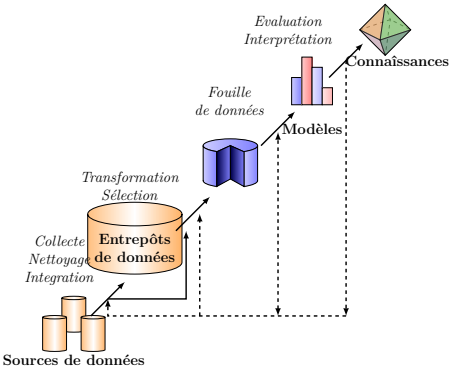


Schéma de l'ECD

6. Utilisation des connaissances découvertes :

- Intégration de ces connaissances dans des systèmes performants,
- Mettre à la disposition des décideurs.

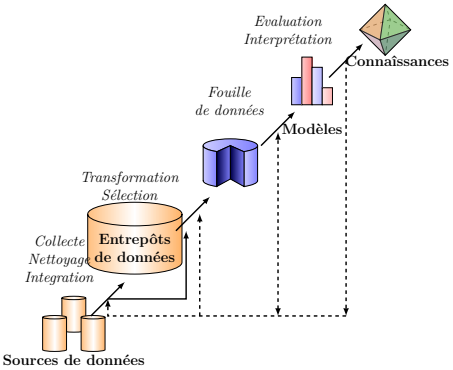
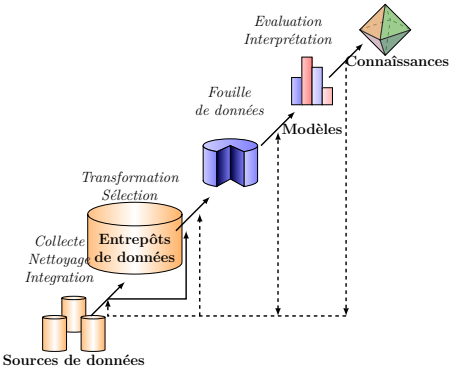


Schéma de l'ECD

6. Utilisation des connaissances découvertes :

- Intégration de ces connaissances dans des systèmes performants,
- Mettre à la disposition des décideurs.



Tâches de fouille de données

Classées en deux catégories

Tâches descriptives

caractérisent les propriétés des données contenues dans un ensemble de données de cibles.

Tâches de fouille de données

Classées en deux catégories

Tâches descriptives

caractérisent les propriétés des données contenues dans un ensemble de données de cibles.

Tâches prédictives

effectuent une induction sur les données actuelles afin de faire des prédictions.

Classification

Classification :

La classification (appelée aussi apprentissage supervisé) est le processus de recherche d'un modèle (ou une fonction) qui décrit et distingue des classes de données ou des concepts.

Classification

Classification :

La classification (appelée aussi apprentissage supervisé) est le processus de recherche d'un modèle (ou une fonction) qui décrit et distingue des classes de données ou des concepts.

- Le modèle est établi en se basant sur l'analyse d'un ensemble de données d'apprentissage.
- il est utilisé pour prédire la classe d'objets dont la classe est inconnue.

Classification

Classification :

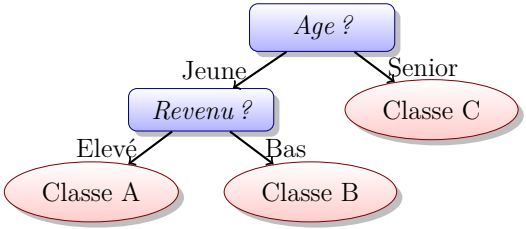
La classification (appelée aussi apprentissage supervisé) est le processus de recherche d'un modèle (ou une fonction) qui décrit et distingue des classes de données ou des concepts.

- Le modèle est établi en se basant sur l'analyse d'un ensemble de données d'apprentissage.
- il est utilisé pour prédire la classe d'objets dont la classe est inconnue.

Arbre de décision

Un arbre de décision est un organigramme ayant une structure arborescente où :

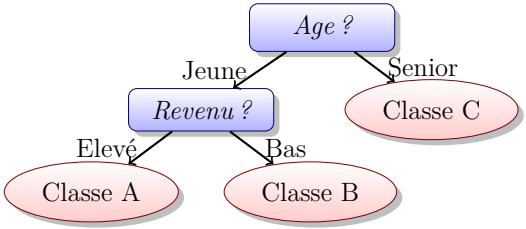
- Chaque noeud représente un test sur une valeur d'attribut,
- Chaque branche représente un résultat de test, et
- Les feuilles représentent des classes.



Arbre de décision

Un arbre de décision est un organigramme ayant une structure arborescente où :

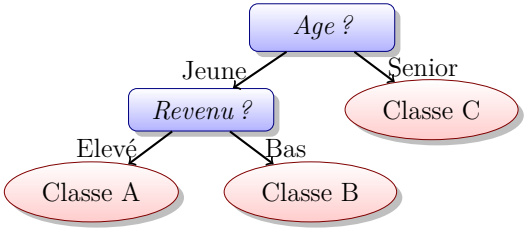
- Chaque noeud représente un test sur une valeur d'attribut,
- Chaque branche représente un résultat de test, et
- Les feuilles représentent des classes.



Arbre de décision

Un arbre de décision est un organigramme ayant une structure arborescente où :

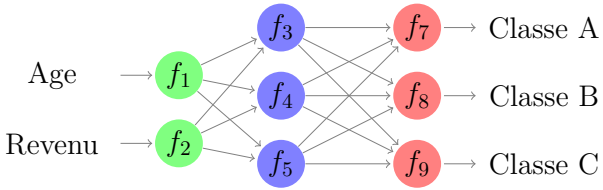
- Chaque noeud représente un test sur une valeur d'attribut,
- Chaque branche représente un résultat de test, et
- Les feuilles représentent des classes.



Réseau de neurones

Un réseau de neurones est généralement une collection de neurones :

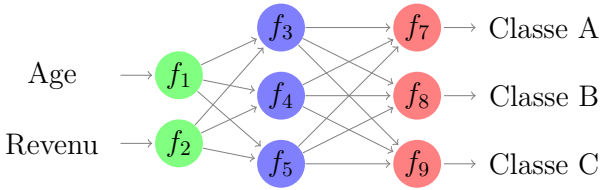
- des unités de traitement,
- des connexions pondérées entre les unités.



Réseau de neurones

Un réseau de neurones est généralement une collection de neurones :

- des unités de traitement,
- des connexions pondérées entre les unités.



Autres classifieurs

- Les modèles probabilistes qui calculent les probabilités pour des hypothèses basées sur le théorème de Bayes.
- Les classifieurs plus proches voisins, qui calculent la distance minimale à partir d'instances ou de prototypes.
- Les règles de classification

$$\begin{array}{ccc} \textbf{Age}(\text{X,'}\textit{Jeune}\text{'}) \text{ ET } \textbf{Revenu}(\text{X,'}\textit{Elevé}\text{'}) & \rightarrow & \textbf{Classe}(\text{X,'}A\text{'}) \\ \textbf{Age}(\text{X,'}\textit{Jeune}\text{'}) \text{ ET } \textbf{Revenu}(\text{X,'}\textit{Bas}\text{'}) & \rightarrow & \textbf{Classe}(\text{X,'}B\text{'}) \\ \textbf{Age}(\text{X,'}\textit{Senior}\text{'}) & \rightarrow & \textbf{Classe}(\text{X,'}C\text{'}) \end{array}$$

Autres classifieurs

- Les modèles probabilistes qui calculent les probabilités pour des hypothèses basées sur le théorème de Bayes.
- Les classifieurs plus proches voisins, qui calculent la distance minimale à partir d'instances ou de prototypes.
- Les règles de classification

$$\begin{array}{ccc} \textbf{Age}(\text{X,'}\textit{Jeune}\text{'}) \text{ ET } \textbf{Revenu}(\text{X,'}\textit{Elevé}\text{'}) & \rightarrow & \textbf{Classe}(\text{X,'}A\text{'}) \\ \textbf{Age}(\text{X,'}\textit{Jeune}\text{'}) \text{ ET } \textbf{Revenu}(\text{X,'}\textit{Bas}\text{'}) & \rightarrow & \textbf{Classe}(\text{X,'}B\text{'}) \\ \textbf{Age}(\text{X,'}\textit{Senior}\text{'}) & \rightarrow & \textbf{Classe}(\text{X,'}C\text{'}) \end{array}$$

Autres classifieurs

- Les modèles probabilistes qui calculent les probabilités pour des hypothèses basées sur le théorème de Bayes.
- Les classifieurs plus proches voisins, qui calculent la distance minimale à partir d'instances ou de prototypes.
- Les règles de classification

$$\begin{array}{ccc} \textbf{Age}(\text{X,'}\textit{Jeune}\text{'}) \text{ ET } \textbf{Revenu}(\text{X,'}\textit{Elevé}\text{'}) & \rightarrow & \textbf{Classe}(\text{X,'}A\text{'}) \\ \textbf{Age}(\text{X,'}\textit{Jeune}\text{'}) \text{ ET } \textbf{Revenu}(\text{X,'}\textit{Bas}\text{'}) & \rightarrow & \textbf{Classe}(\text{X,'}B\text{'}) \\ \textbf{Age}(\text{X,'}\textit{Senior}\text{'}) & \rightarrow & \textbf{Classe}(\text{X,'}C\text{'}) \end{array}$$

Exemples d'application de la classification

- Identification de signature des documents sensibles (*correspondance, aucune correspondance*).
- Identification d'empreinte digitale numérique dans des applications de sécurité (*correspondance, aucune correspondance*).
- Attribuer un crédit bancaire considérant de la qualité de la clientèle, et les possibilités financières (*bon, moyen, mauvais*).
- L'efficacité du traitement d'un médicament en présence d'un ensemble de maladies symptômes (*bon, moyen, mauvais*).

Exemples d'application de la classification

- Identification de signature des documents sensibles (*correspondance, aucune correspondance*).
- Identification d'empreinte digitale numérique dans des applications de sécurité (*correspondance, aucune correspondance*).
- Attribuer un crédit bancaire considérant de la qualité de la clientèle, et les possibilités financières (*bon, moyen, mauvais*).
- L'efficacité du traitement d'un médicament en présence d'un ensemble de maladies symptômes (*bon, moyen, mauvais*).

Exemples d'application de la classification

- Identification de signature des documents sensibles (*correspondance, aucune correspondance*).
- Identification d'empreinte digitale numérique dans des applications de sécurité (*correspondance, aucune correspondance*).
- Attribuer un crédit bancaire considérant de la qualité de la clientèle, et les possibilités financières (*bon, moyen, mauvais*).
- L'efficacité du traitement d'un médicament en présence d'un ensemble de maladies symptômes (*bon, moyen, mauvais*).

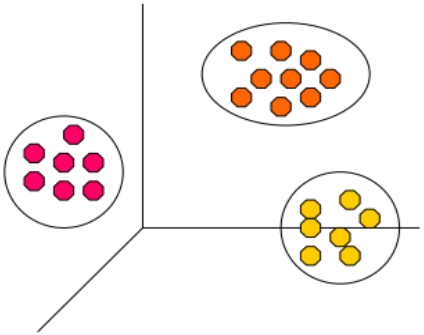
Exemples d'application de la classification

- Identification de signature des documents sensibles (*correspondance, aucune correspondance*).
- Identification d'empreinte digitale numérique dans des applications de sécurité (*correspondance, aucune correspondance*).
- Attribuer un crédit bancaire considérant de la qualité de la clientèle, et les possibilités financières (*bon, moyen, mauvais*).
- L'efficacité du traitement d'un médicament en présence d'un ensemble de maladies symptômes (*bon, moyen, mauvais*).

Segmentation(Clustering)

La segmentation se rapporte à la catégorisation d'un ensemble d'objets de données dans des clusters.

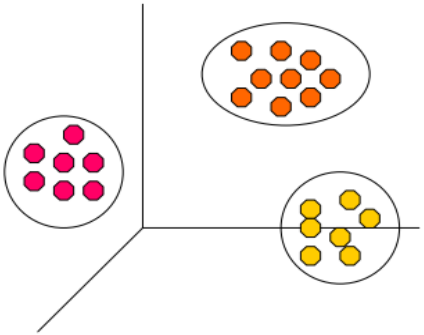
- Elle est aussi appelée classification non supervisée.
- Un cluster est une collection d'objets de données :
  - Similaires les uns aux autres dans le même segment.
  - Différents des objets dans d'autres segments.



Segmentation(Clustering)

La segmentation se rapporte à la catégorisation d'un ensemble d'objets de données dans des clusters.

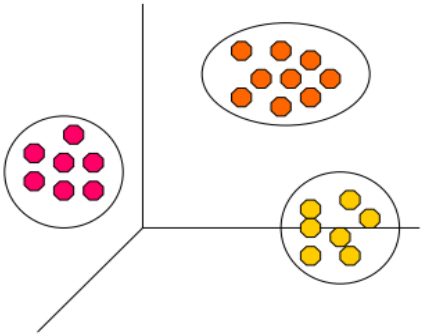
- Elle est aussi appelée classification non supervisée.
- Un cluster est une collection d'objets de données :
  - Similaires les uns aux autres dans le même segment,
  - Différents des objets dans d'autres segments.



Segmentation(Clustering)

La segmentation se rapporte à la catégorisation d'un ensemble d'objets de données dans des clusters.

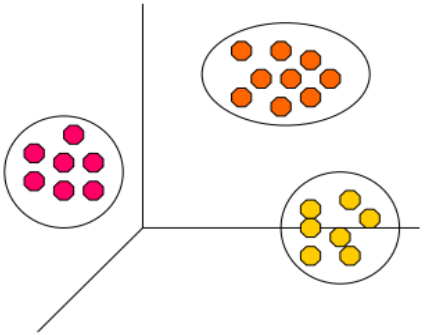
- Elle est aussi appelée classification non supervisée.
- Un cluster est une collection d'objets de données :
  - Similaires les uns aux autres dans le même segment,
  - Différents des objets dans d'autres segments.



Segmentation(Clustering)

La segmentation se rapporte à la catégorisation d'un ensemble d'objets de données dans des clusters.

- Elle est aussi appelée classification non supervisée.
- Un cluster est une collection d'objets de données :
  - Similaires les uns aux autres dans le même segment,
  - Différents des objets dans d'autres segments.

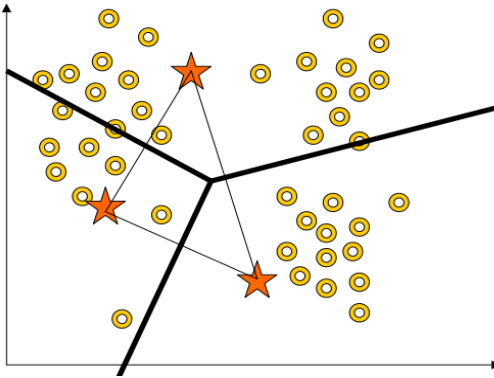


Approches de clustering

- **Méthode de partitionnement :**
  - Créer un partitionnement initial.
  - Utiliser une stratégie de contrôle itérative pour l'optimiser.

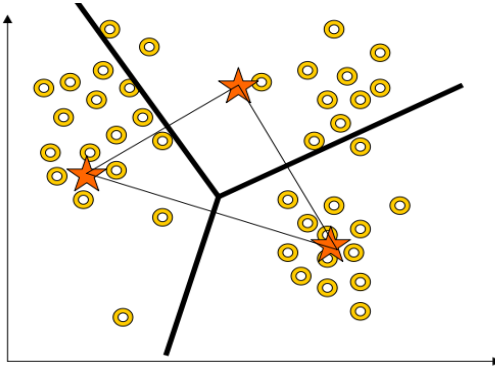
Approches de clustering

- **Méthode de partitionnement :**
  - Créer un partitionnement initial.
  - Utiliser une stratégie de contrôle itérative pour l'optimiser.



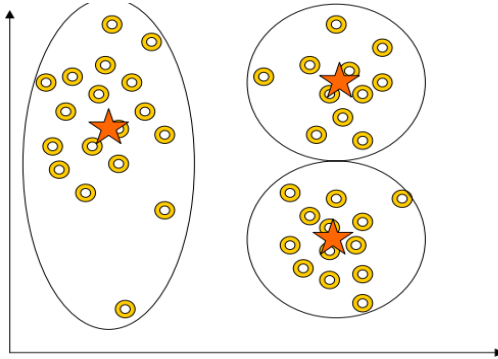
Approches de clustering

- **Méthode de partitionnement :**
  - Créer un partitionnement initial.
  - Utiliser une stratégie de contrôle itérative pour l'optimiser.



Approches de clustering

- **Méthode de partitionnement :**
  - Créer un partitionnement initial.
  - Utiliser une stratégie de contrôle itérative pour l'optimiser.

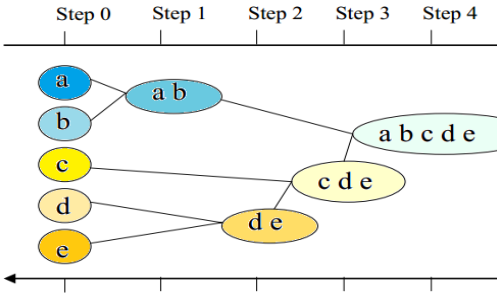


Approches de clustering

- **Méthode de partitionnement :**
  - Créer un partitionnement initial.
  - Utiliser une stratégie de contrôle itérative pour l'optimiser.

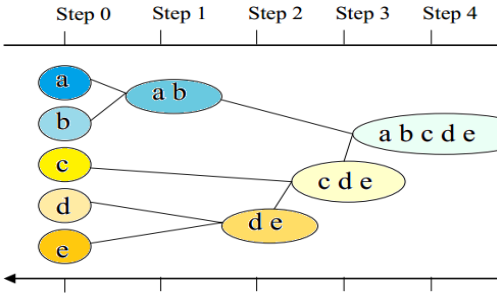
Approches de clustering

- **Méthodes hiérarchiques :**
  - Construire une hiérarchie de clusters (appelé dendrogramme),
  - Non seulement un partitionnement unique des objets.
  - Utiliser une condition de terminaison. (ex. Nombre de clusters).
  - Méthodes basées sur la densité : utiliser les fonctions de densité de voisinage.



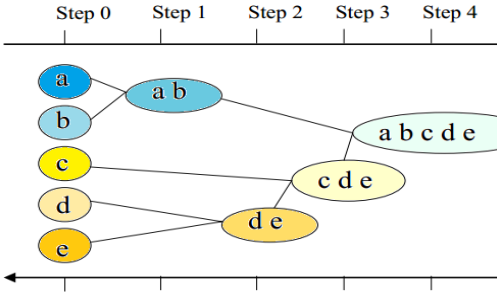
Approches de clustering

- **Méthodes hiérarchiques :**
  - Construire une hiérarchie de clusters (appelé dendrogramme),
  - Non seulement un partitionnement unique des objets.
  - Utiliser une condition de terminaison. (ex. Nombre de clusters).
  - Méthodes basées sur la densité : utiliser les fonctions de densité de voisinage.



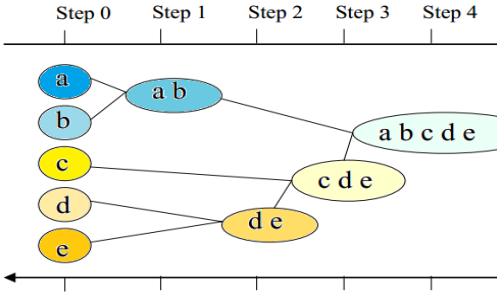
Approches de clustering

- **Méthodes hiérarchiques :**
  - Construire une hiérarchie de clusters (appelé dendrogramme),
  - Non seulement un partitionnement unique des objets.
  - Utiliser une condition de terminaison. (ex. Nombre de clusters).
  - Méthodes basées sur la densité : utiliser les fonctions de densité de voisinage.



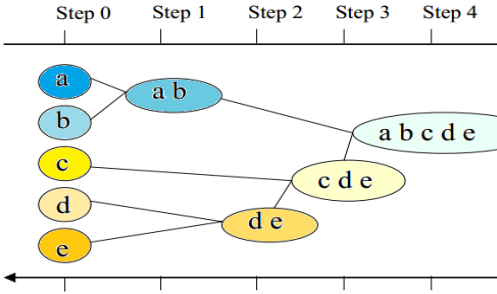
Approches de clustering

- **Méthodes hiérarchiques :**
  - Construire une hiérarchie de clusters (appelé dendrogramme),
  - Non seulement un partitionnement unique des objets.
  - Utiliser une condition de terminaison. (ex. Nombre de clusters).
  - Méthodes basées sur la densité : utiliser les fonctions de densité de voisinage.



Approches de clustering

- **Méthodes hiérarchiques :**
  - Construire une hiérarchie de clusters (appelé dendrogramme),
  - Non seulement un partitionnement unique des objets.
  - Utiliser une condition de terminaison. (ex. Nombre de clusters).
  - Méthodes basées sur la densité : utiliser les fonctions de densité de voisinage.



Exemples d'application de la segmentation

- La reconnaissance de formes et le traitement d'images.
- Analyse des données spatiales : créer des cartes thématiques dans les systèmes d'information géographique (SIG).
- Bioinformatique : la détermination des groupes de signatures à partir d'une base de données de gènes.
- Web : clustering des fichiers log pour découvrir des modèles d'accès similaires.

Exemples d'application de la segmentation

- La reconnaissance de formes et le traitement d'images.
- Analyse des données spatiales : créer des cartes thématiques dans les systèmes d'information géographique (SIG).
- Bioinformatique : la détermination des groupes de signatures à partir d'une base de données de gènes.
- Web : clustering des fichiers log pour découvrir des modèles d'accès similaires.

Exemples d'application de la segmentation

- La reconnaissance de formes et le traitement d'images.
- Analyse des données spatiales : créer des cartes thématiques dans les systèmes d'information géographique (SIG).
- Bioinformatique : la détermination des groupes de signatures à partir d'une base de données de gènes.
- Web : clustering des fichiers log pour découvrir des modèles d'accès similaires.

Exemples d'application de la segmentation

- La reconnaissance de formes et le traitement d'images.
- Analyse des données spatiales : créer des cartes thématiques dans les systèmes d'information géographique (SIG).
- Bioinformatique : la détermination des groupes de signatures à partir d'une base de données de gènes.
- Web : clustering des fichiers log pour découvrir des modèles d'accès similaires.

Règles d'association

La fouille de règles d'association se rapporte à la découverte des relations entre les attributs d'un ensemble de données appelé souvent ensemble des transactions.

- Une transaction est l'ensemble des articles achetés ensemble par les clients.
- Une règle est normalement exprimée sous la forme $A \Rightarrow B$, où A et B sont des ensembles d'attributs de l'ensemble de données. Cela implique que les transactions qui contiennent A contiennent B avec une grande probabilité.
- La règle peut s'écrire sous une autre forme :

SI $<$*certaines conditions satisfaites*$>$ ALORS $<$*prédire les valeurs pour certains autres attributs*$>$,

Règles d'association

La fouille de règles d'association se rapporte à la découverte des relations entre les attributs d'un ensemble de données appelé souvent ensemble des transactions.

- Une transaction est l'ensemble des articles achetés ensemble par les clients.
- Une règle est normalement exprimée sous la forme $A \Rightarrow B$, où A et B sont des ensembles d'attributs de l'ensemble de données. Cela implique que les transactions qui contiennent A contiennent B avec une grande probabilité.
- La règle peut s'écrire sous une autre forme :

SI <*certaines conditions satisfaites* > ALORS <*prédire les valeurs pour certains autres attributs*>,

Règles d'association

La fouille de règles d'association se rapporte à la découverte des relations entre les attributs d'un ensemble de données appelé souvent ensemble des transactions.

- Une transaction est l'ensemble des articles achetés ensemble par les clients.
- Une règle est normalement exprimée sous la forme $A \Rightarrow B$, où A et B sont des ensembles d'attributs de l'ensemble de données. Cela implique que les transactions qui contiennent A contiennent B avec une grande probabilité.
- La règle peut s'écrire sous une autre forme :

SI $<$*certaines conditions satisfaites* $>$ ALORS $<$*prédire les valeurs pour certains autres attributs*$>$,

Règles d'association

Une règle d'association $A \Rightarrow B$ peut être identifié lorsque le support et la confiance de la règle sont largement supérieurs aux seuils respectifs.

- Le support de la règle d'association est le rapport entre le nombre de transactions contenant à la fois A et B sur le nombre total de transactions dans la base de données.
- La confiance de la règle d'association est la proportion du nombre de transactions contenant à la fois A et B sur le nombre total de transactions contenant A.

Règles d'association

Une règle d'association $A \Rightarrow B$ peut être identifié lorsque le support et la confiance de la règle sont largement supérieurs aux seuils respectifs.

- Le support de la règle d'association est le rapport entre le nombre de transactions contenant à la fois A et B sur le nombre total de transactions dans la base de données.
- La confiance de la règle d'association est la proportion du nombre de transactions contenant à la fois A et B sur le nombre total de transactions contenant A.

Règles d'association

Par exemple, la règle :

$$\text{Age}(X, 20..29) \wedge \text{revenu}(X, 40000..49000) \Rightarrow \text{achète}(X\ ,\ \textit{“Ordinatur portable”})$$
$$(\text{ support } 2\%, \text{ confiance } 60\%)$$

signifie que :

- 2% des clients sont âgés de 20 à 29 ans ayant un revenu compris entre 40.000 et 49.000 et ont achetés un ordinateur portable.
- Il y a une probabilité de 60% qu'un client dans cet intervalle d'âge et de revenu va acheter un ordinateur portable.

Règles d'association

Par exemple, la règle :

$$\text{Age}(X, 20..29) \wedge \text{revenu}(X, 40000..49000) \Rightarrow \text{achète}(X\ ,\ \textit{“Ordinatur portable”})$$
$$(\text{ support } 2\%, \text{ confiance } 60\%)$$

signifie que :

- 2% des clients sont âgés de 20 à 29 ans ayant un revenu compris entre 40.000 et 49.000 et ont achetés un ordinateur portable.
- Il y a une probabilité de 60% qu'un client dans cet intervalle d'âge et de revenu va acheter un ordinateur portable.