Clustering

Marie Beurton-Aimar

November 24, 2020

Le clustering

- Qu'est-ce que le clustering?
- Types de données en clustering: variables d'intervalles, variables binaires, variables ordinales ou nominales ou ratio, variables mixtes
- Catégories de clustering
- Méthodes de partitionnement: k-means, k-médoïdes
- Méthodes hiérarchiques
- Méthodes basées sur la densité
- Méthodes basées sur les grilles
- Méthodes basées sur un modèle
- Analyse d'exception

Qu'est-ce que le clustering?

- Analyse de clustering
  - regroupement d'objets similaires en clusters
- Un cluster: une collection d'objets
  - similaires au sein d'un même cluster
  - dissimilaires aux objets appartenant à d'autres clusters
- Classification non supervisée
  - pas de classe prédéfinies
- Applications typiques
  - afin de mieux comprendre les données
  - comme prétraitement avant d'autres analyses

Applications générales

- Reconnaissance de motifs
- Analyse de données spatiales
  - détection de clusters géographiques et leur compréhension
- Traitement d'image
- Economie
- www
  - classification de documents
  - analyse de log: motifs / séquences d'accès
- Bio-informatique
  - données d'expression
  - réseau d'intéraction
  - familles de gènes, protéines

Qu'est-ce qu'un bon clustering?

- Une bonne méthode va produire des clusters dont les éléments ont:
  - une forte similarité intra-classe
  - une faible similarité inter-classe
- La qualité d'un clustering dépend de la mesure de similarité
- La qualité d'une méthode peut aussi être mesurée par sa capacité à trouver quelques ou tous les motifs intéressants

Prérequis

- Mise à l'échelle
- Capacité à gérer différents tyes d'attributs
- Découverte de clusters avec des formes arbitraires
- Besoin minimum de connaissances du domaine pour déterminer les paramètres
- Capacité à gérer le bruit et les exceptions
- Indifférent à l'ordre des données en entrée
- Nombre de dimensions
- Incorporation de contraintes par l'utilisateur
- Interprétabilité et utilisabilité

Structure de données

- Matrice de données

$$\begin{bmatrix} x_{11} & \dots & x_{1f} & \dots & x_{1p} \\ \dots & \dots & \dots & \dots & \dots \\ x_{i1} & \dots & x_{if} & \dots & x_{ip} \\ \dots & \dots & \dots & \dots & \dots \\ x_{n1} & \dots & x_{nf} & \dots & x_{np} \end{bmatrix}$$

- Matrice de distance (ou dissimilarité)

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \dots & \dots & 0 \end{bmatrix}$$

Similarité et dissimilarité

- Métrique de similarité/dissimilarité: exprimée en terme d'une fonction de distance, typiquement $d(i, j)$
- Fonction de distance dépend du type des données: binaire, catégoriel, ordinal ou continu
- Pondération des dimensions selon l'application et la sémantique des données
- Difficulté de définir "suffisamment similaires"
  - la réponse est très subjective

Types de données

- continu sur un intervalle
  - ex: poids, taille
- binaire
- nominal
  - ex: couleur
- ordinal
- échelle variable
  - ex: croissance optimale des bactéries, durée de la radioactivité
- mixte

Valeurs continues sur un intervalle
Standardisation

- Standardiser les données : s'affranchir des unités de mesures
- Ecart absolu à la moyenne

$$s_f = \frac{1}{n}(\mid x_{1f} - m_f \mid + \mid x_{2f} - m_f \mid + \ldots + \mid x_{nf} - m_f \mid)$$

- Calculer la mesure standardisée (z-score)

$$z_{if} = \frac{x_{if} - m_f}{s_f}$$

Valeurs continues sur un intervalle
Fonction de distance

- Distance de Minkowski:

$$d(i,j) = \sqrt[q]{(\mid x_{i1} - x_{j1} \mid^q + \mid x_{i2} - x_{j2} \mid^q + \ldots + \mid x_{ip} - x_{jp} \mid^q)}$$

avec $i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ et $j = (x_{j1}, x_{j2}, \ldots, x_{jp})$, deux objets à p dimensions et q un entier positif

- si $q = 1$: distance de Manhattan

$$d(i,j) = (\mid x_{i1} - x_{j1} \mid + \mid x_{i2} - x_{j2} \mid + \ldots + \mid x_{ip} - x_{jp} \mid$$

- si $q = 2$: distance Euclidienne

$$d(i,j) = \sqrt[2]{(\mid x_{i1} - x_{j1} \mid^2 + \mid x_{i2} - x_{j2} \mid^2 + \ldots + \mid x_{ip} - x_{jp} \mid^2)}$$

- Propriétés
  - $d(i,i) = 0$
  - $d(i,j) \geq 0$ (positive)
  - $d(i,j) = d(j,i)$ (fonction symétrique)
  - $d(i,j) \leq d(i, k + d(k,j))$ (inégalité triangulaire)

Valeurs binaires

- table de contingence

		objet j	
		1	0
objet i	1	a	b
	0	c	d

- coefficient simple d'appariement (invariant, si la variable est symétrique)

$$d(i,j) = \frac{b+c}{a+b+c+d}$$

- coefficient de Jaccard (non invariant, si la variable est asymétrique)

$$d(i,j) = \frac{b+c}{a+b+c}$$

Dissimilarité de valeurs binaires

- exemple:

Nom	Genre	Fièvre	Toux	Test1	Test2	Test3	Test4
Jacques	M	O	N	P	N	N	N
Marie	F	O	N	P	N	P	N
Jean	M	O	P	N	N	N	N

- `Genre` est symétrique
- les autres sont asymétriques
- soit O et P $= 1$, et N $= 1$

$$d(Jacques, Marie) = \frac{0+1}{2+0+1} = 0.33$$
$$d(Jacques, Jean) = \frac{1+1}{1+1+1} = 0.67$$
$$d(Jean, Marie) = \frac{1+2}{1+1+2} = 0.75$$

Variables nominales

- Généralisation des valeurs binaires: plus de 2 états
- Méthode 1: appariement simple
  - m: nbr d'appariements, p: nbr total de variables

$$d(i,j) = \frac{p-m}{p}$$

- Méthode 2: utiliser un grand nombre de variables binaires
  - création d'une variable binaire pour chacun des états d'une variable nominale

Variable ordinale

- L'ordre est important: rang
- La variable peut être traitée comme une variable continue sur un intervalle
  - remplacer x_{if} par son rang $r_{if} \in \{1, \ldots, M_f\}$
  - transformer chaque variable sur $[0, 1]$ en remplaçant le rang r_{if} de l'objet i pour la varaible f

$$z_{if} = \frac{r_{if} - 1}{M_f - 1}$$

  - calculer la dissimilarité en utilisant les méthodes de valeurs continues sur un intervalle

A échelle variable

- Mesure positive sur une échelle non linénaire, approximativement une échelle exponentielle telle que Ae^{BT} ou Ae^{-BT}
- Méthodes:
  - les traiter comme des variables continues sur un intervalle: mauvais choix
  - appliquer une transformation logarithmique puis les traiter commes des variables continues sur un intervalle

$$y_{if} = \log(x_{if})$$

  - les traiter commes des variables ordinales en traitant leur rang

Variables de types mixte

- Les objets peuvent être décrits avec tous les types de données
  - binaire, symétrique, binaire asymétrique, nominale, ordinale, ...
- Utilisation d'une formule pondérée pour combiner leurs effets

$$d(i,j) = \frac{\sum_{f=1}^{p} \delta_{ij(f)} d_{ij}(f)}{\sum_{f=1}^{p} \delta_{ij}(f)}$$

Principales approches

- Partitionnement
  - partitionnement des objets et évalue les partitions
- Hiérarchie
  - décomposition hiérarchique d'ensembles d'objets
- Densité
  - basée sur une fonction de densité ou de connectivité
- Grille
  - basée sur une structure de granularité à plusieurs niveaux
- Basée sur un modèle
  - construction d'un modèle pour chaque cluster

Partitionnement

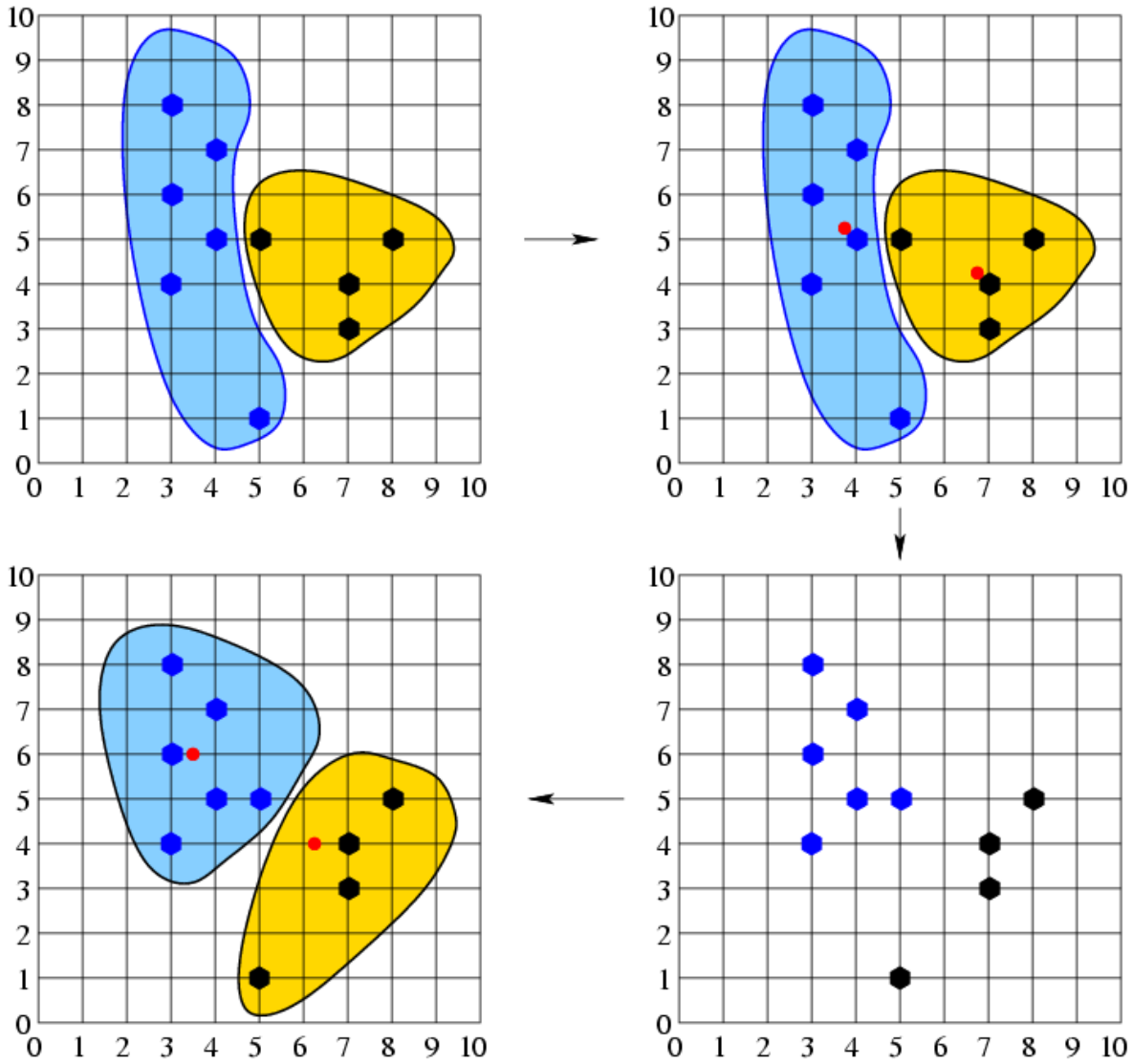
- Construire une partition de la base de données D contenant n objets en un ensemble de k clusters
- Entant donné k, trouver une partition en k clusters qui optimisnet le crière de partitionnement
  - Optimum global: traiter toutes les partitions exhaustivement
  - Heuristique: `k-means` ou `k-médoïdes`
    - `k-means`: chaque cluster est représenté par son centre
    - `k-médoïdes` ou `PAM` (Partition Around Medoids): chaque cluster est représenté par l'objet le plus proche du centre du cluster

k-means

4 étapes:

1. Partitionner les objets en k ensembles non vides
2. Calculer le centroïde de chaque partition / cluster
3. Assigner à chaque objet le cluster dont le centroïde est le plus proche
4. Boucler en 2, jusqu'à ce que les clusters soient stables

k-means: exemple



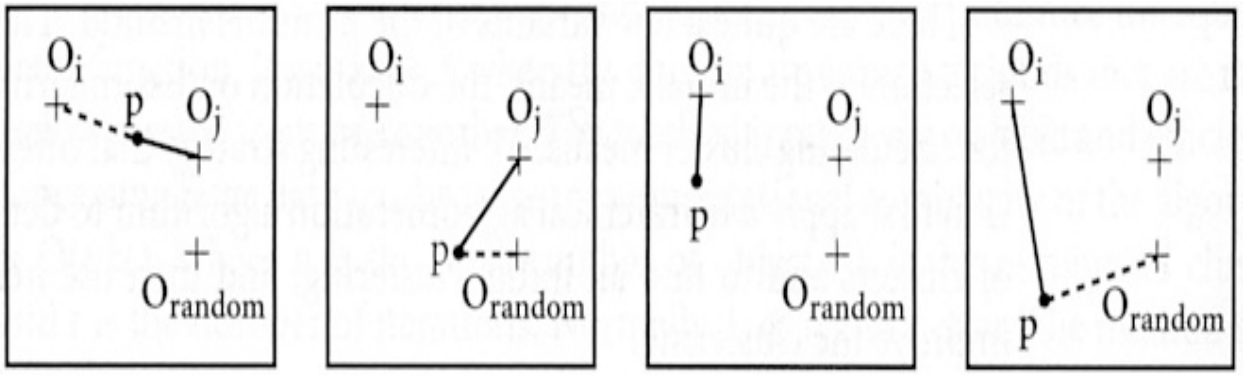
k-means: remarques

- Avantage:
  - relativement effice: $O(tkn)$, avec n le nbr d'objets, t le nbr d'itérations et en général t et $k << n$
- Faiblesses:
  - Utilisable seulement lorsque la moyenne est définie. Que faire dans le cas de donnéess nominales?
  - Besoin de spécifier k à l'avance
  - Ne gère pas le bruit et les exceptions
  - Ne trouve que des clusters de forme convexe

k-medoïdes

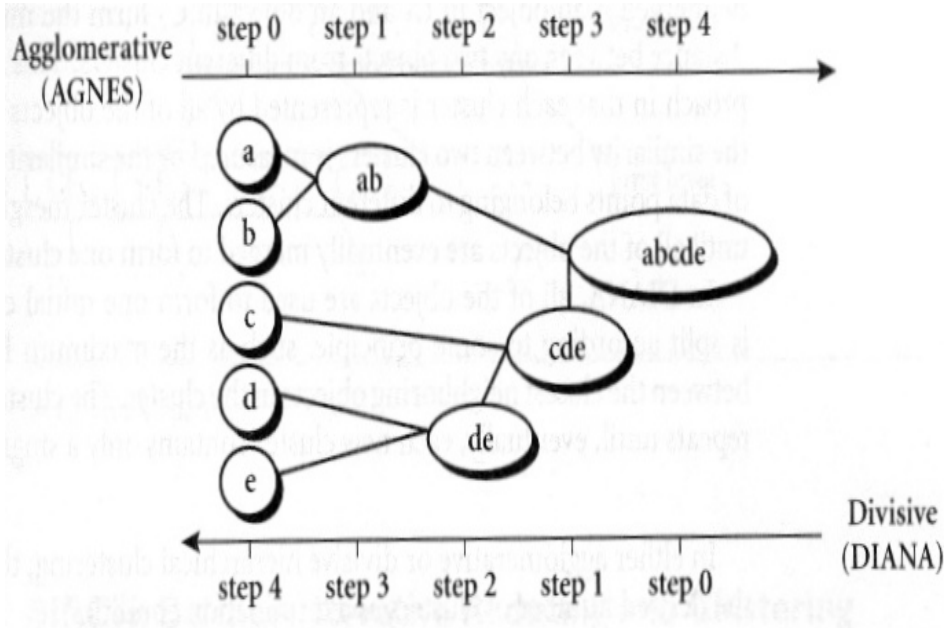
- Trouve des représentants, appelés médoïdes, dans les clusters
- PAM (Partitioning Around Medoids)
  - médoïde: l'objet d'un cluster pour lequel la distance moyenne à tous les autres objets du cluster est minimale
  - critère d'erreur: $E = \sum_{i=1}^{k} \sum_{p \in C_i} d(p, m)^2$
- Algorithme
  1. Sélectionner k objets arbitrairement
  2. Assigner le reste des objets au plus proche médoïde
  3. Sélectionner un objet non médoïde et échanger si le critère d'erreur peut être réduit
  4. Répéter 2 et 3 jusqu'à ne plus pouvoir réduire le critère d'erreur

PAM: exemple


O_i
p
O_j
O_{random}
1. Reassigned to O_i
2. Reassigned to O_{random}
3. No change
4. Reassigned to O_{random}
• data object
+ cluster center
—before swapping
---after swapping

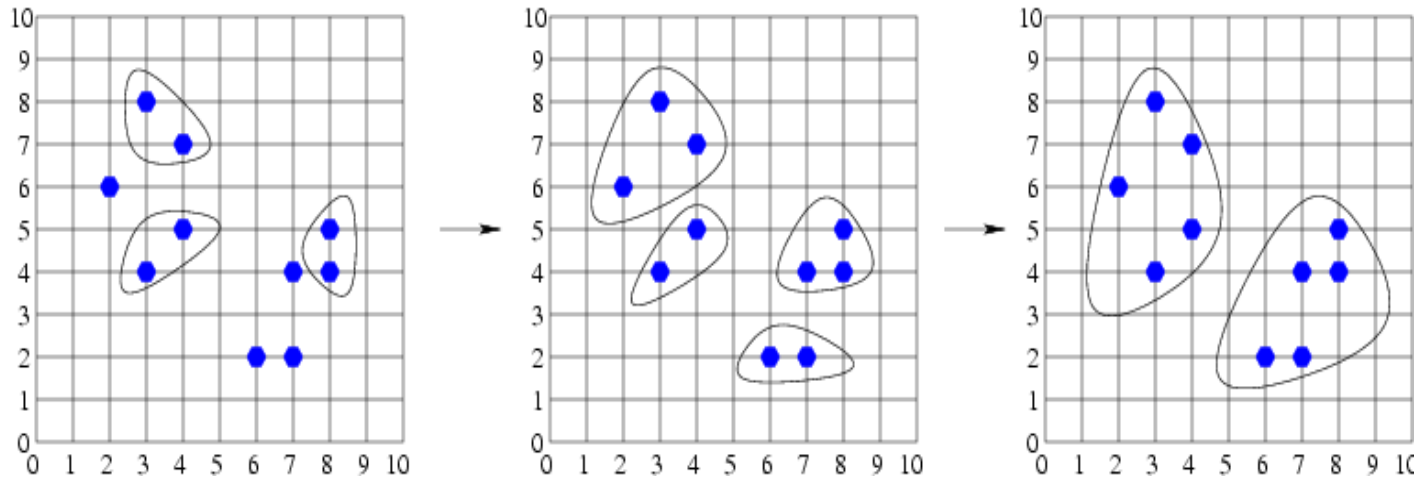
Clustering hiérarchique

- Utilisation d'une matrice de distance: ne nécessite pas de spécifier le nombre de clusters



AGNES (Agglomerative Nesting)

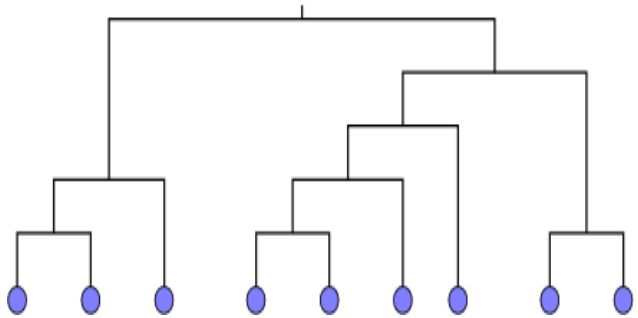
- Utilise une matrice de dissimilarité
- Fusionne les nœuds les moins dissimilaires



Dendogramme

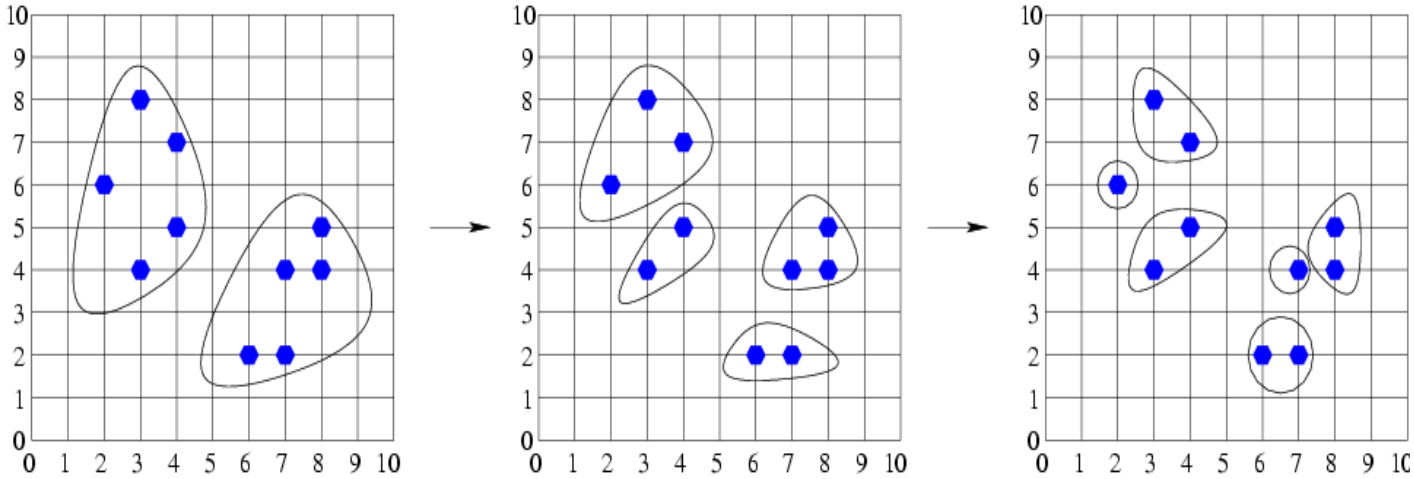
Un dendogramme illustre comment les clusters sont fusionnés hiérarchiquement

- Décompose les données en plusieurs niveaux imbriqués de partitionnement
- Un clustering est obtenu en coupant le dendogramme au niveau choisi



DIANA (Divisive Analysis)

C'est l'inverse d'AGNES

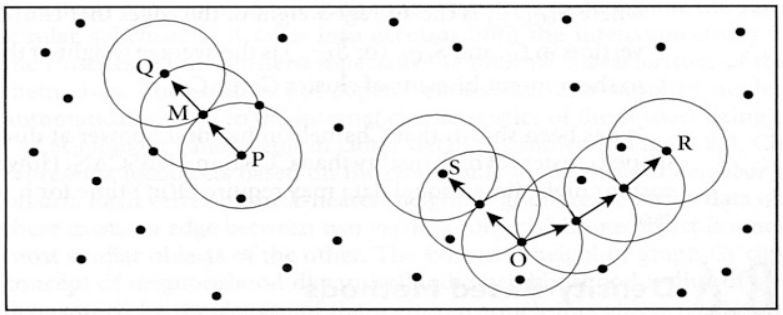


Méthodes basées sur la densité)

- Principales caractéristiques
  - Cluster de forme arbitraire
  - Gestion du bruit
  - Besoin d'un paramètre de densité comme critère d'arrêt
- 2 paramètres:
  - Eps: rayon maximal de voisinage
  - MinPts: nbr minimal de points dans le voisinage défini par Eps

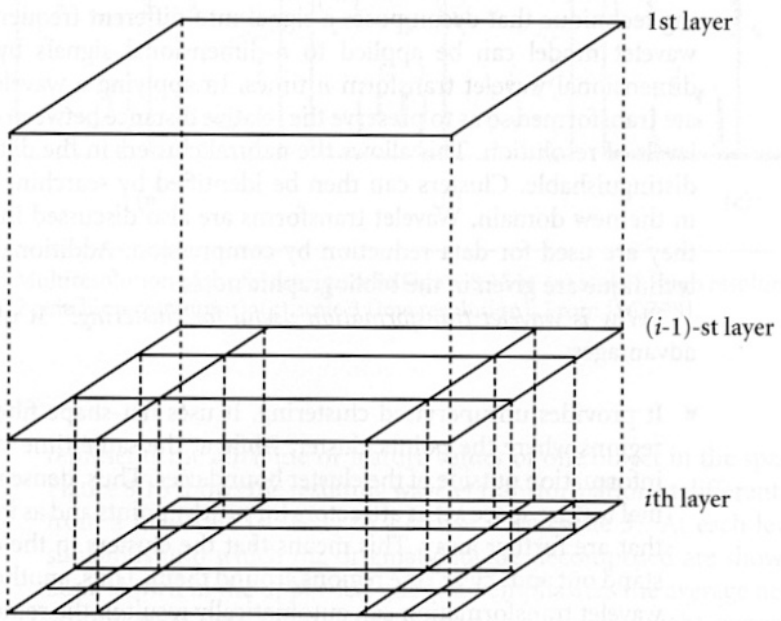
Méthodes basées sur la densité

- $N_{Eps}(Q) : \{M \in D \mid dist(Q, M) \leq Eps\}$
- Un point Q est **directement atteignable** depuis un point M si
  - Q appartient à $N_{Eps}(M)$
  - $\mid N_{Eps}(M) \mid \geq Min_{Eps}$ (ici: MinPts=3)
- Un point Q est **atteignable** d'un point P si
  - il existe une chaîne de points $M_1, \ldots, M_n$ telle que $M_1 = P$ et $M_n = Q$ et que les M_{i+1} sont directement atteignables des M_i
- Un point S est **connecté** à un point R si
  - il existe un point O tel que S et R sont atteignables depuis O



Méthodes basées sur une grille

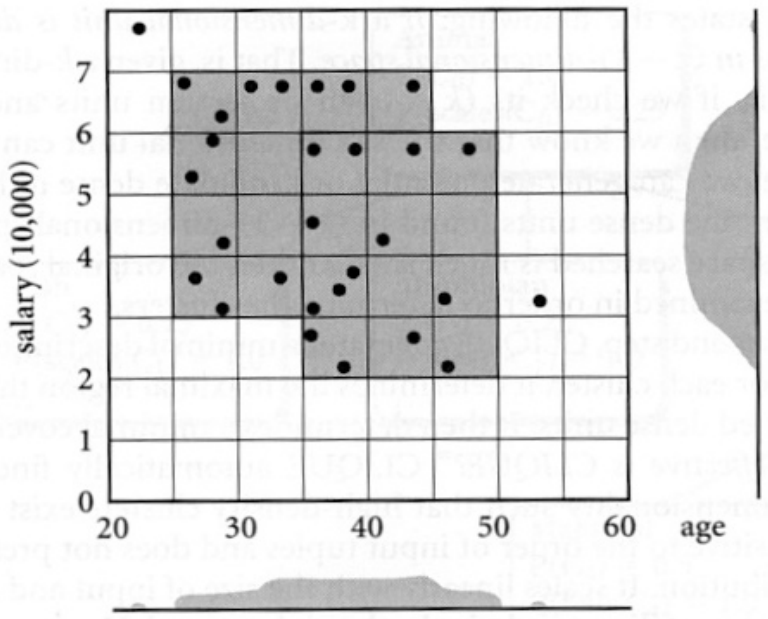
- Utilisation d'une grille à des résolutions multiples comme structure de données
- L'espace est divisé en cellules rectangulaires

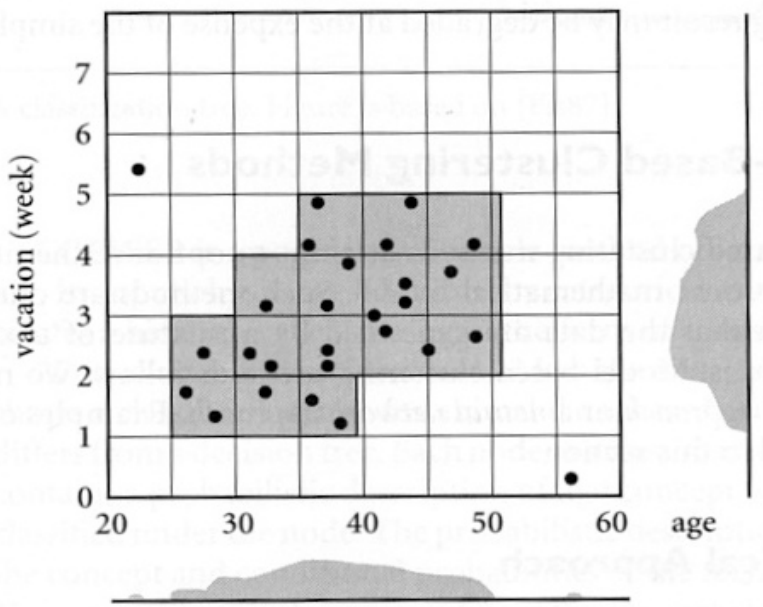


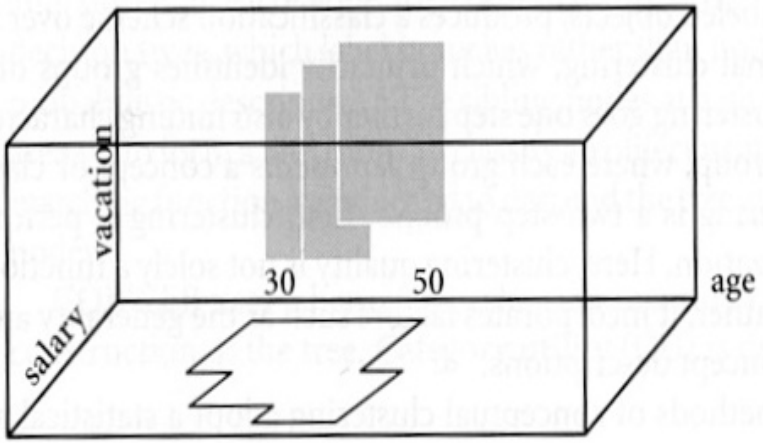
Méthodes basées sur une grille

- Chaque cellule de niveau i est divisée en un certain nombre de cellules plus petites au niveau i+1
- Informations statistiques calculées et stockées à chaque niveau
- Approche descendante
- Suppression des cellules non pertinentes pour les itérations suivantes
- Répéter le processus jusqu'à atteindre le niveau plus bas
- Avantages:
  - parallélisable, mise à jour incrémentale
  - $O(k)$, où k est le nombre de cellules au plus bas niveau
- Faiblesses
  - les bords des clusters sont soit horizontaux, soit verticaux, pas de diagonale!

Méthodes basées sur une grille

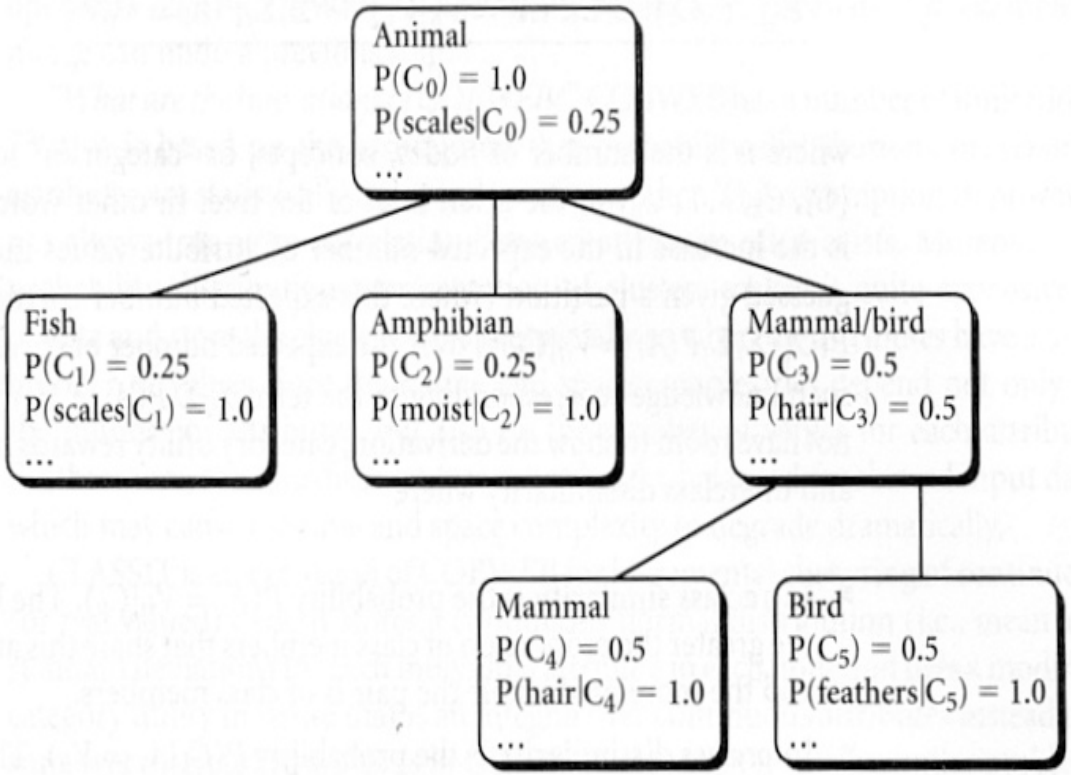






Méthodes basées sur un modèle

- Optimisation d'un modèle mathématique par rapport aux données
- Approches statistiques et intelligence artificielle



Analyse des exceptions

- Exceptions
  - ensemble d'objets qui sont particulièrement dissimilaires au reste des données
- Problème
  - trouver les n objets les plus exceptionnels
- Approches
  - statistique
  - distance
  - déviation

Approches statistiques

- Suppose que la distribution des données suit un modèle
  - ex: distribution normale
    cf. figure
- Utilisation de test de discordance
  - distribution des données
  - distribution de paramètres (moyenne, variance, . . .)
  - nombre attendu d'exceptions
- Faiblesses
  - la plupart des tests est pour un attribut seul
  - dans bien des cas, la distribution n'est pas connue

Distance

- Afin de pallier les défauts des approches statistiques
  - besoin d'analyses multidimensionnelles sans connaissance de la distribution des données
- Exception basée sur la distance
  - $ebd(p, d)$ est un objet O d'un ensemble de données E tel qu'au moins une partie p des objets de E est à une distance supérieure à d de O

Déviation

- Identification d'une exception en examinant les caratéristiques principales des objets d'un groupe
- Les objets "déviant" de cette description sont considérés comme des exceptions
- Technique séquentielle
  - simule la manière dont les humains distinguent les objets inhabituels d'une série d'objets supposés ressemblant
- Technique OLAP
  - utilisation d'un cube de données pour identifier des régions contenant des anomalies dans un grand espace multidimensionnel

Exemple OLAP

Sum of sales	*Month*											
	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Total		1%	−1%	0%	1%	3%	−1%	−9%	−1%	2%	−4%	3%

Avg. sales	*Month*											
Item	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Sony b/w printer		9%	−8%	2%	−5%	14%	−4%	0%	41%	−13%	−15%	−11%
Sony color printer		0%	0%	3%	2%	4%	−10%	−13%	0%	4%	−6%	4%
HP b/w printer		−2%	1%	2%	3%	8%	0%	−12%	−9%	3%	−3%	6%
HP color printer		0%	0%	−2%	1%	0%	−1%	−7%	−2%	1%	−4%	1%
IBM desktop computer		1%	−2%	−1%	−1%	3%	3%	−10%	4%	1%	−4%	−1%
IBM laptop computer		0%	0%	−1%	3%	4%	2%	−10%	−2%	0%	−9%	3%
Toshiba desktop computer		−2%	−5%	1%	1%	−1%	1%	5%	−3%	−5%	−1%	−1%
Toshiba laptop computer		1%	0%	3%	0%	−2%	−2%	−5%	3%	2%	−1%	0%
Logitech mouse		3%	−2%	−1%	0%	4%	6%	−11%	2%	1%	−4%	0%
Ergo-way mouse		0%	0%	2%	3%	1%	−2%	−2%	−5%	0%	−5%	8%

Avg. sales	*Month*											
Region	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
North		−1%	−3%	−1%	0%	3%	4%	−7%	1%	0%	−3%	−3%
South		−1%	1%	−9%	6%	−1%	−39%	9%	−34%	4%	1%	7%