Thème 4 : Clustering

Partie 2: Détermination du nombre de clusters; Clustering des données catégoriques et transactionelles; Évaluation des résultats

S. Wang

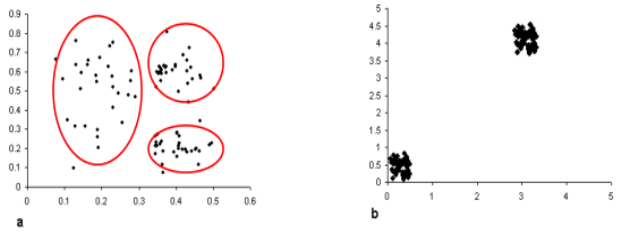
Objectifs

- Identification du nombre de clusters pour K-means et FCM
- Clustering des données catégoriques
- Clustering des données transactionnelles
- Évaluation des résultats des algorithmes de clustering

Identification automatique du nombre de clusters

Position du problème

➢ *Difficulté de détecter s'il existe une structure en clusters.*

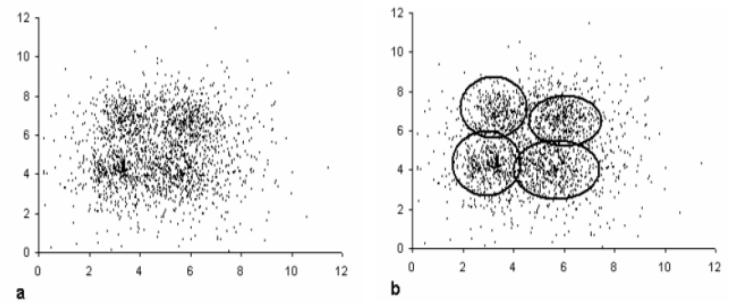


Existe-t-il une structure de cluster?

Identification automatique du nombre de clusters

Position du problème

➢ *Difficulté de déterminer le nombre exacte de clusters.*



Identification automatique du nombre de clusters

Position du problème

➢ La caractéristique principale d'un algorithme de clustering, c'est qu'il fonctionne de manière non supervisée.

➢ Comme toute approche non supervisée, la qualité de ces résultats doit être validée.

➢ Valider le résultat d'un algorithme de clustering implique immédiatement la recherche du bon nombre de clusters.

➢ Par "le bon nombre de clusters" on veut dire que chacun de ces clusters doivent faire apparaître une structure sous-jacente aux données et ainsi permettre de faciliter leur interprétation. Chaque cluster doit avoir une signification pour l'expert du domaine.

Identification automatique du nombre de clusters

Position du problème

Typiquement deux situations se présentent:

– **Trop de clusters**: cette situation peut entraîner une grande confusion car certains clusters sont "artificiels", c'est-à-dire qu'ils ne représentent aucune réalité du domaine concerné.

– **Pas assez de clusters**: cet autre cas peut cacher des aspects importants présents dans les données. Par exemples, on peut séparer un ensemble de patients en deux groupes : les patients sains et malades. Mais il peut être plus intéressant pour le médecin d'utiliser une structure en trois clusters faisant ressortir les patients sains, malades et à risque.

Position du problème

A ce problème, l'approche la plus utilisée consiste à :

1) Exécuter les algorithmes de clustering avec différents nombres de clusters.
2) Evaluer leurs résultats et ce à partir d'une comparaison entre ces derniers.

➢L'évaluation des résultats est basée essentiellement sur l'utilisation des **indices de validité (cluster validity index).**

Indice de validité?

Formellement, un indice de validité est une fonction/critère qui mesure la qualité du résultat final d'un algorithme de clustering. En général, un indice de validité est définie pour réaliser un équilibre entre la cohésion interne des clusters et la séparation entre les clusters.

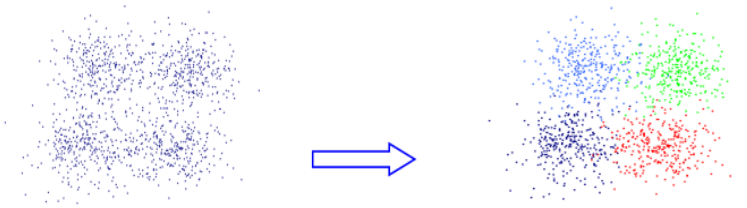
La détermination automatique du nombre de clusters se fait par l'utilisation d'un processus itératif, qui consiste à exécuter un algorithme de clustering avec différents nombres de clusters, afin de trouver le nombre de clusters qui optimise (la plus petite ou la plus grande valeur) l'indice de validité en question.

Stratégie d'implémentation (« Model Selection »)

Entrée : $X = \{x_1, \ldots x_n\}$ ensemble de données, C_{min} le nombre minimum de clusters, C_{max} le nombre maximum de clusters.
Sortie : le nombre de cluster c qui optimise un indice de validité.

1. Pour $K = C_{min}$, jusqu'à C_{max} ;
   a- Appliquez l'algorithme de clustering.
   b- Calculer la valeur de l'indice de validité.
2. Calculer c_f de tel sort que l'indice $V_d(K)$ soit optimal.

Exemple



Combien de clusters ?

4 clusters

Choix des valeur de Cmin et Cmax

➢ Généralement Cmin = 2

➢ Pour le choix de Cmax il n'y a aucune règle formelle.
Intuitivement : Cmax << n (n le nombre de données)
permis les choix possible:

$$C_{max} = \sqrt{n}$$

La cohésion interne « compctness » et l'isolation externe « séparation » sont deux facteurs importants pour évaluer la validité du résultat d'un algorithme de clustering.

La cohésion interne est une mesure dédiée à évaluer la variation interne manifesté par la concentration des objets appartenant au même cluster autour du centre

L'isolation externe : maximiser la distance entre les points représentant les clusters

➢ Tous les indices proposés tentent à de formaliser un compromis entre l'isolation externe et la cohésion interne.

Indices de validité

Types d'indices de validité

En pratique, on distingue deux types de partition : dure " Hard " et floue " Fuzzy ", d'où la nécessité d'utiliser des indices de validité conformes à ces différents types de partitions. Il y a deux types d'indices de validité :

– 1) les indices dédiés à évaluer la validité des partitions "hard". Dans ce contexte, nous focalisons sur les indices de validité qui sont utilisés avec K-means

– 2) les indices dédiés à évaluer la validité des partitions floues. Dans ce contexte, nous focalisons sur les indices de validité qui sont utilisés avec FCM.

Indices de validités

Indices de validité pour l'algorithme K-means

- Plusieurs indices de validité ont été proposés.
- Exemple d'indice de validité

$$V_{BW} = \frac{dispersion\ \text{inter} - clusters}{dispersion\ \text{intra} - clusters} = \frac{B}{W}$$

$$B = \frac{1}{K}\sum_{c=1}^{K}\|v_c - \bar{v}\|^2 \qquad W = \frac{1}{K}\sum_{c=1}^{K}\frac{1}{n_c}\sum_{x_i \in cj}\|x_i - v_c\|^2$$

K: est le nombre de clusters

v_c: le centre de cluster c (c = 1, …, K)

$\bar{v}$: le centre global (vecteur des moyenne prises sur toutes les n données)

- **Choisir K qui maximise V_{BW}**

❑ Question: Pourquoi on choisi la valeur de K qui maximise V_{BW}?

Indices de validités

Indices de validité pour l'algorithme FCM

$$V_{SC}(U,V,X) = \frac{Sep}{Comp} \qquad \textit{à maximiser}$$

$$Sep = trace(S_B) \qquad Comp = \sum_{c=1}^{K} trace(\Sigma_C)$$

S_B est la matrice de séparation floue

$$S_B = \sum_{c=1}^{K}\sum_{i=1}^{n} u_{ic}^m (v_c - \bar{v})(v_c - \bar{v})^T$$

$\sum_c$ est la matrice de covariance floue du cluster c

$$\sum_c = \frac{\sum_{i=1}^{n} u_{ic}^m (x_i - v_c)(x_i - v_c)^T}{\sum_{i=1}^{n} u_{ic}^m}$$

Plan

Clustering des données catégorique

Exemple de données catégorique

	director	actor	genre	
t_1 (Godfather II)	Scorsese	De Niro	Crime	Cluster 1
t_2 (Good Fellas)	Coppola	De Niro	Crime	
t_3 (Vertigo)	Hitchcock	Stewart	Thriller	Cluster 2
t_4 (N by NW)	Hitchcock	Grant	Thriller	
t_5 (Bishop's Wife)	Koster	Grant	Comedy	Cluster 3
t_6 (Harvey)	Koster	Stewart	Comedy	

Algorithme *K*-representatives

Papier:

O.M. San, V-N. Huynh and Y. Nakamori

An Alternative Extension of the *K*-means Algorithm for Clustering Categorical Data

International Journal of Applied Mathematics and Computer Science, vol. 14, No 2, pp. 214-247, 2004.

K-representatives

Notation

• **Données en entrée :** Un ensemble d'objets $X = \{x_1, x_2, \ldots, x_n\}$ dans l'espace de dimension *d*, *n* le nombre d'objets.

$x_i = (x_{i1}, x_{i2}, \ldots, x_{id})$ représente le $i^{ème}$ objet; et x_{ij} correspond à la valeur du $j^{ième}$ attribut pour le $i^{ème}$ objet.

• **Sortie:** Un ensemble de cluster C_l $|l=\{1, \ldots, K\}$; K représente le nombre de cluster

• Soit D_j le domaine des valeurs « catégorique » pour une dimension j

Exemple

Soit l'ensemble de données suivant :

A A Y

B A Y

F A A

D1 = {A, B, F}

D2 = {A}

D3 = {Y, A}

K-representatives

Notation

Soit Q = (q1, q2, …, q_d) : Le représentant du cluster C (cluster representative) – *(équivalent au centre de cluster dans le contexte des données numériques)*

Avec $q_j = \{(c_j, f_{c_j}) | c_j \in D_j\} \quad (j = 1,...,d)$

f_{c_j} La fréquence de la catégorie c_j dans le cluster C

Pour un cluster C: $f_{c_j} = n_{cj} / p$

avec p : le nombre d'objet dans cluster C

K-representatives

Notation

Exemple sur l'estimation du Q

Soit le cluster suivant qui contient 3 objets représentés dans l'espace de 4 dimensions

A B D E
A B D B
A C A F

Q = {q1, q2, q3, q4}

q1 = {(A, 1)}

q2 = {(B, 0.66), (C, 0.33)}

q3 = {(D, 0.66), (A, 0.33)}

q4 = {(E, 0.33), (B, 0.33), (F, 0.33)}

K-representatives

Mesure de dissemblance (dissimilarity measure)

La dissemblance entre un objet *x* et Q est définie comme suit :
Équivalent à la distance d'un objet vers un centre de cluster dans le contexte des données numériques

$$d(x,Q) = \sum_{j=1}^{d} \sum_{c_j \in D_j} f_{c_j} \times \delta(x_j, c_j)$$

avec

$$\delta(x_j, c_j) = \begin{cases} 0 \; si \; x_j = c_j \\ 1 \; si \; x_j \neq c_j \end{cases}$$

K-representatives

Mesure de dissemblance -- Simplification

$$d(x,Q) = \sum_{j=1}^{d} \sum_{c_j \in D_j} f_{c_j} \times \delta(x_j, c_j)$$

$$= \sum_{j=1}^{d} \sum_{(c_j \in D_j, c_j \neq x_j)} f_{c_j}$$

$$= \sum_{j=1}^{d} (1 - f_{x_j})$$

f_{x_j} La fréquence de la catégorie x_j dans le cluster C

K-representatives -- Algorithme

K-representatives tente de minimiser la fonction suivante $E = \sum_{l=1}^{K} \sum_{x_i \in Cl} d(x_i, Q_l)$

L'algorithme *K*-representatives

Entrée : Un ensemble d'objets X ={x1,x2,…, xn}, nombre de cluster *K*
Sortie : C = {C_1, C_2, …, C_K}

1. Diviser ALÉATOIREMENT l'ensemble de données X en K clusters;
2. Pour chaque cluster C_l (l = 1, …, K), calculer son représentant Q_l
3. Pour chaque objet xi (i = 1, …, n) calculer d(x_i, Q_l)
4. Réassigner x_i au cluster C_l, tel que d(x_i, Q_l) est la plus petite
5. Répéter les étapes 2, 3 et 4 jusqu'à la stabilité de la partition

Remarque

en générale K-representatives à les mêmes avantage et faiblisse de l'algorithme K-means

Evaluation d'un algorithme de clustering

➢ Techniques d'évaluation externe qui visent à mesurer la correspondance entre la partition identifiée par un algorithme de clustering et la partition originale.

➢ Ces techniques sont utilisées seulement lorsque la partition originale est connue par l'utilisateur.

Matrice de confusion

➢ La matrice de confusion est un outil servant à mesurer la qualité d'un système de clustering et de classification.

➢ Les colonnes de la matrice de confusion représentent la répartition des points dans les classes réelles (répartition originale des points)

➢ Les lignes représentent la répartition des points dans les clusters identifiés par un algorithme de clustering

Matrice de confusion

Exemple 1 (adopté à partir d'un exemple dans Wikipédia)

On considère un algorithme de clustering dont le but est de regrouper du courrier électronique en deux clusters : courriels **normaux** et courriels **spam**. On va vouloir savoir combien de courriels normaux seront faussement estimés comme du spam et combien de spams ne seront pas estimés comme tels. On va supposer qu'on a testé notre algorithme de clustering avec 100 courriels normaux et 100 courriels de spam. Les résultats de notre algorithme de clustering sont représentés dans la matrice de confusion suivante :

		Clusters réelles	
		normal	spam
Clusters identifiées	normal	95	3
	spam	5	97

Matrice de confusion

Exemple 1

		Clusters réelles	
		normal	spam
Clusters identifiées	normal	95	3
	spam	5	97

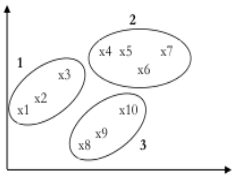
La matrice se lit comme suit :

- Sur les 100 courriels normaux, 95 sont estimés comme tels et 5 sont estimés comme du spam;
- Sur les 100 spams, 3 sont estimés comme courriels normaux, et 97 sont estimés comme du spam;
- Sur les 98 courriels que l'algorithme a estimé comme normaux, 3 sont en fait du spam;
- Sur les 102 courriels que l'algorithme a estimé comme spam, 5 sont en fait des courriels normaux.

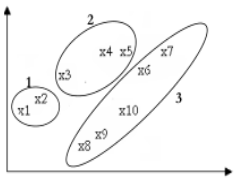
Matrice de confusion

Exemple 2

Partitionnement original



Partitionnement identifier par un algorithme de clustering



Matrice de confusion

Exemple 2

Partitionnement original

x1	1
x2	1
x3	1
x4	2
x5	2
x6	2
x7	2
x8	3
x9	3
x10	3

Résultat d'un alog. de clustering

x1	1
x2	1
x3	2
x4	2
x5	2
x6	3
x7	3
x8	3
x9	3
x10	3

Matrice de confusion

		Partitionnement original		
		c1	c2	c3
Résultat	C1	2	0	0
	C2	1	2	0
	C3	0	2	3

Matrice de confusion

➢ **Pour un algorithme de clustering qui produit un bon résultat** → Chaque ligne de la matrice de confusion doit avoir une cellule qui contient un plus grand nombre de point comparativement aux autres cellules (de la même ligne).

➢ **Pour un algorithme de clustering qui produit un mauvais résultat** → les points sont distribués de façons aléatoires dans la matrice.

Remarque

➢ Lorsque la matrice de confusion est très grande (c.à.d. le nombre de clusters est très grand), une lecture simple de cette matrice ne suffit pas pour juger la qualité du résultat.

➢ Pour cette raison il y a des d'autres mesures, qui sont estimées à partir de la matrice de confusion, pour évaluer la qualité des résultats.

Erreur de clustering

L'erreur de clustering est estimée selon la formule suivante:

$$ER = \frac{\sum_{i}\sum_{j \neq h} n_{ij}}{n}$$

n_{ij} : les différentes valeurs dans la matrice de confusion

h: l'index du cluster original c_j avec une valeur maximum n_{ij}

➢ La valeur de ER est toujours entre 0 et 1.

➢ Une valeur de ER proche de zéro indique un bon clustering.

Exemple

		Partitionnement original		
		c1	c2	c3
Résultat	C1	2	0	0
	C2	1	2	0
	C3	0	2	3

$$ER = \frac{(0+0)+(1+0)+(2+0)}{10} = \frac{3}{10} = 0.3$$

Problème avec l'erreur de clustering

Exemple

Partitionnement original

x1	1
x2	1
x3	1
x4	2
x5	2
x6	2
x7	2
x8	3
x9	3
x10	3

Résultat d'un alog. de clustering

x1	1
x2	1
x3	2
x4	3
x5	3
x6	4
x7	4
x8	5
x9	5
x10	6

Matrice de confusion

Résultat	Partitionnement original c1	c2	c3
C1	2	0	0
C2	1	0	0
C3	0	2	0
C4	0	2	0
C5	0	0	2
C6	0	0	1

Erreur de clustering

ER = 0

➢Dans ce cas, l'algorithme à réussi à identifier des clusters « pure » (c.à.d. l'algorithme ne mélange pas les points qui appartiennent au même cluster original avec d'autres clusters)

➢ Cependant, l'algorithme ne réussi pas à identifier les « vrais » clusters selon la répartition originale des points

Table de contingence

➢ Il y a d'autres mesures, plus solide, qui se basent sur la table de contingence. Cette dernière est elle-même estimé à partir de la matrice de confusion.

Format de la table de contingence

Partition identifier : R	Partition originale : O	
	N_{11}	N_{10}
	N_{01}	N_{00}

N_{11} : le nombre de paires d'objets qui sont dans le même cluster dans O et R.
N_{10} : le nombre de paires d'objets qui sont dans le même cluster dans O mais pas dans R.
N_{01} : le nombre de paires d'objets qui sont dans le même cluster dans R mais pas dans O.
N_{00} : le nombre de paires d'objets qui sont dans différents clusters dans O et dans R.

Table de contingence

Soit

$n_{i.}$: la somme des éléments de la ligne i de la matrice de confusion.
$n_{.j}$: la somme des éléments de la colonne j de la matrice de confusion.
n_{ij} : les différents éléments de la matrice de confusion.

Exemple

Partition identifié	Partition originale					somme
	n_{11}	n_{12}	n_{13}	…	n_{1K}	$n_{1.}$
	n_{21}	n_{22}	n_{13}	…	n_{2K}	$n_{2.}$
	⋮	⋮	⋮		⋮	⋮
	n_{c1}	n_{c2}	n_{c3}	…	n_{cK}	$n_{c.}$
Somme	$n_{.1}$	$n_{.2}$	$n_{.3}$		$n_{.K}$	

Table de contingence

$$N_{11} = \sum_i \sum_j \binom{n_{ij}}{2}$$

$$N_{10} = \sum_j \binom{n_{.j}}{2} - \left[\sum_i \sum_j \binom{n_{ij}}{2}\right] = \sum_j \binom{n_{.j}}{2} - N_{11}$$

$$N_{01} = \sum_i \binom{n_{i.}}{2} - \left[\sum_i \sum_j \binom{n_{ij}}{2}\right] = \sum_i \binom{n_{i.}}{2} - N_{11}$$

n : la taille de l'ensemble de données

$$N_{11} + N_{10} + N_{01} + N_{00} = \binom{n}{2} \Longrightarrow N_{00} = \binom{n}{2} - (N_{11} + N_{10} + N_{01})$$

i varie de 1 jusqu'à c | c représente le nombre de lignes de la matrice de confusion

j varie de 1 jusqu'à K | K représente le nombre de colonnes de la matrice de confusion

Rappel

$$\binom{\alpha}{2} = \frac{\alpha!}{2! \times (\alpha-2)!} = \frac{\alpha \times (\alpha-1) \times (\alpha-2)!}{2 \times (\alpha-2)!} = \frac{\alpha \times (\alpha-1)}{2}$$

Quelques mesures de qualité

$$Jaccard = \frac{N_{11}}{N_{01} + N_{10} + N_{11}}$$

$$Fowlkes = \frac{N_{11}}{\sqrt{(N_{11} + N_{01}) \times (N_{11} + N_{10})}}$$

$$Rand = \frac{N_{11} + N_{00}}{N_{01} + N_{10} + N_{11} + N_{00}}$$

Quelques mesures de qualité

$$\text{Pr}esion = \frac{N_{11}}{N_{11} + N_{01}} \qquad \text{Re}call = \frac{N_{11}}{N_{11} + N_{10}}$$

$$F - measure = \frac{1}{\frac{1}{2}\left(\frac{1}{Precision} + \frac{1}{Recall}\right)} = \frac{2 * Precision * Recall}{Precison + Recall}$$

➢ La valeur de toutes ces mesures est toujours entre 0 et 1.
➢Une valeur proche de 1 indique un clustering de bonne qualité (la similarité entre la partition originale O et la partition identifiée par un algorithme de clustering R est relativement élevée).