1

Clustering (Segmentation)

2

Problématique

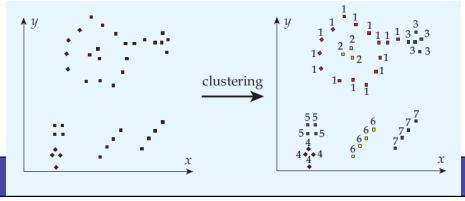
Objectif = structuration des données

Clustering (en anglais) = Classification (en français) non supervisée

On cherche à regrouper les points proches en groupe ou classes ou cluster

Les points ou objets différents appartiennent à des groupes différents

Les classes peuvent être assez bien définies



3

Problématique

Les classes peuvent aussi être assez imbriquées,

avoir des formes bizarres,

ou pire

Mais surtout ne pas être en 2 dimensions

4

Problématique

Soient N instances de données à k attributs,

Trouver un partitionnement en c clusters (groupes) ayant un sens (Similitude)

Affectation automatique de “labels” aux clusters

c peut être donné, ou “découvert”

Plus difficile que la classification car les classes ne sont pas connues à l’avance (non supervisé)

Attributs

- Numériques (distance bien définie)
- Enumératifs ou mixtes (distance difficile à définir)

5

Exemples d’applications

Marketing : segmentation du marché en découvrant des groupes de clients distincts à partir de bases de données d’achats.

Environnement : identification des zones terrestres similaires (en termes d’utilisation) dans une base de données d’observation de la terre.

Assurance: identification de groupes d’assurés distincts associés à un nombre important de déclarations.

Planification de villes : identification de groupes d’habitations suivant le type d’habitation, valeur, localisation géographique, …

Médecine : Localisation de tumeurs dans le cerveau

- Nuage de points du cerveau fournis par le neurologue
- Identification des points définissant une tumeur

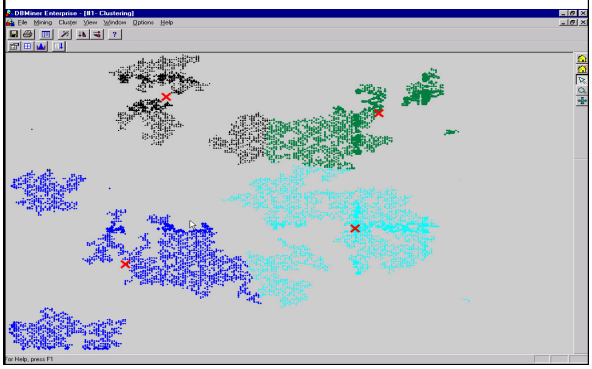
6

Domaine d’application

Domaine	Forme des données	Clusters
Text mining	Textes, Mails	Textes proches, Dossiers automatiques
Web mining	Textes et images	Pages web proches
Bioinformatique	Gènes	Gènes ressemblants
Marketing	Infos clients, produits achetés	Segmentation de la clientèle
Segmentation d’image	Images	Zones homogènes dans l’image
Web log analysis	Clickstream	Profils utilisateurs

Exemple: segmentation de marchés

7



Qualité d'un clustering

8

Une bonne méthode de clustering produira des clusters d'excellente qualité avec :

- Similarité intra-classe importante
- Similarité inter-classe faible

La qualité d'un clustering dépend de :

- La mesure de similarité utilisée
- L'implémentation de la mesure de similarité

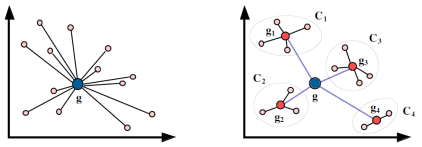
La qualité d'une méthode de clustering est évaluée par son habilité à découvrir certains ou tous les "patterns" cachés.

Objectifs du clustering

9

Maximiser les distances inter-clusters

Minimiser les distances intra-cluster

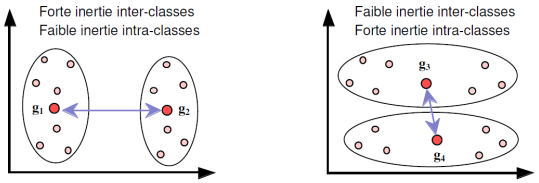


Inertie totale des points = Inertie Intra + Inter

Bisson 2001

Objectifs du clustering

10



Approches de Clustering

11

Algorithmes de Partitionnement: Construire plusieurs partitions puis les évaluer selon certains critères

Algorithmes hiérarchiques: Créer une décomposition hiérarchique des objets selon certains critères

Algorithmes basés sur la densité: basés sur des notions de connectivité et de densité

Algorithmes de grille: basés sur un structure à multi-niveaux de granularité

Algorithmes à modèles: Un modèle est supposé pour chaque cluster ensuite vérifier chaque modèle sur chaque groupe pour choisir le meilleur

Méthodes de clustering - Caractéristiques

12

Extensibilité

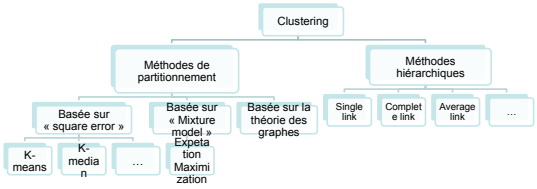
Habilité à traiter différents types de données

Découverte de clusters de différents formes

Connaissances requises (paramètres de l'algorithme)

Habilité à traiter les données bruitées et isolées.

Taxonomie



Notions de base nécessaires au regroupement d'individus

Commune à toutes les méthodes

- Mesure de distance/dissimilarité ou de proximité/similarité inter-individus

Méthodes non-hierarchiques : critère de qualité d'un regroupement

Méthode hierarchique : critère d'agrégation ou de division

Attention : cela va déterminer le type de groupe à identifier

Notion de proximité

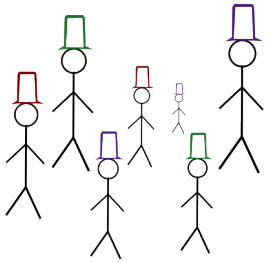
Vocabulaire

Mesure de dissimilarité (DM) : plus la mesure est faible plus les points sont similaires (~ distance)

Mesure de similarité (SM) : plus la mesure est grande, plus les points sont similaires

DM = borne - SM

Mesure de la similarité



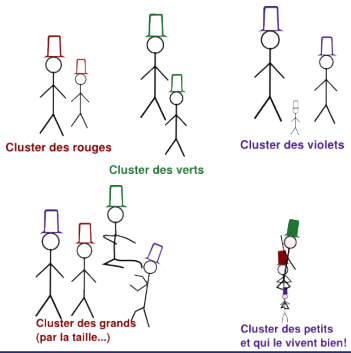
Il n'y a pas de définition unique de la similarité entre objets

- Différentes mesures de distances d(x,y)

La définition de la similarité entre objets dépend de :

- Le type des données considérées
- Le type de similarité recherchée

Mesure de similarité



Distance

Propriétés d'une distance :

1. $d(x,y) \geq 0$
2. $d(x,y) = 0$ iff $x = y$
3. $d(x,y) = d(y,x)$
4. $d(x,z) \leq d(x,y) + d(y,z)$

Similarié : vérifie s(I,j)=s(j,i), s(i,j) >= 0; s(i,i)>=s(i,j)

Distance – Données numériques

19

Combiner les distances : Soient x=(x1,…,xn) et y=(y1, …,yn)

Exemples numériques :

Distance euclidienne : $d(x,y)=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2}$

Distance de Manhattan : $d(x,y)=\sum_{i=1}^{n}|x_i-y_i|$

Distance de Minkowski : $d(x,y)=\sqrt[q]{\sum_{i=1}^{n}|x_i-y_i|^q}$

q=1 : distance de Manhattan.

q=2 : distance euclidienne

Distance données énumératives

20

Champs discrets :

- Données binaires : d(0,0)=d(1,1)=0, d(0,1)=d(1,0)=1
- Donnée énumératives : distance nulle si les valeurs sont égales et 1 sinon.
- Donnée énumératives ordonnées : idem. On peut définir une distance utilisant la relation d'ordre.

Distance – Données binaires

21

Table de concordances/discordances

		Individu *j* 1	0	*sum*
	1	*a*	*b*	*a+b*
Individu *i*	0	*c*	*d*	*c+d*
	sum	*a+c*	*b+d*	*m*

a= nombre de positions où i vaut 1 et j vaut 1

• Exemple oi=(1,1,0,1,0) et oj=(1,0,0,0,1)
a=1, b=2, c=1, d=2

Distance – Données binaires

22

- **% de concordance (distance euclidienne) :** $d(i,j)=\frac{a+d}{m}$
- **Double pondération des concordances**

$$d(i,j)=\frac{2(a+d)}{2(a+d)+b+c}$$

- **Coefficient de correspondance simple** (similarité invariante, si la variable binaire est **symétrique**) : $d(i,j)=\frac{b+c}{m}$
- **Coefficient de Jaccard** (similarité non invariante, si la variable binaire est **asymétrique**): $d(i,j)=\frac{b+c}{a+b+c}$

Variables binaires (I)

23

Variable symétrique: Ex. le sexe d'une personne, i.e coder masculin par 1 et féminin par 0 c'est pareil que le codage inverse

Variable asymétrique: Ex. Test HIV. Le test peut être positif ou négatif (0 ou 1) mais il y a une valeur qui sera plus présente que l'autre. Généralement, on code par 1 la modalité la moins fréquente

- 2 personnes ayant la valeur 1 pour le test sont plus similaires que 2 personnes ayant 0 pour le test

Distance – Données binaires

24

Exemple : dissimilarité entre variables binaires

- Table de patients

Nom	Sexe	Fièvre	Toux	Test-1	Test-2	Test-3	Test-4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	P	N	N	N	N

- 8 attributs, avec
  - Sexe un attribut symétrique, et
  - Les attributs restants sont asymétriques
  - (test VIH, …)

Distance – Données binaires

25

Les valeurs Y et P sont initialisées à 1, et la valeur N à 0.

Calculer la distance entre patients, basée sur le coefficient de Jaccard.

Jack	M	1	0	1	0	0	0
Mary	F	1	0	1	0	1	0
Jim	M	1	1	0	0	0	0

Jack/Mary	1	0
1	A=2	B=0
0	C=1	D=3

Jack/Jim	1	0
1	A=1	B=1
0	C=1	D=3

Jim, Mary	1	0
1	A=1	B=1
0	C=2	D=2

$$d(jack, mary) = \frac{0+1}{2+0+1} = 0.33$$

$$d(jack, jim) = \frac{1+1}{1+1+1} = 0.67$$

$$d(jim, mary) = \frac{1+2}{1+1+2} = 0.75$$

Distance – Données énumératives

26

Généralisation des variables binaires, avec plus de 2 états, e.g., rouge, jaune, bleu, vert

Méthode 1: correpondance simple

- m: # de correspondances, p: # total de variables

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

Variables Ordinales

27

Une variable ordinale peut être discrète ou continue

L'ordre peut être important, ex: classement

Peuvent être traitées comme les variables intervalles

- remplacer x_{if} par son rang $r_{if} \in \{1, \ldots, M_f\}$
- Remplacer le rang de chaque variable par une valeur dans [0, 1] en remplaçant la variable f dans l'objet I par

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

- Utiliser une distance pour calculer la similarité

Données mixtes

28

Soit - transformation des variables numériques en variables catégorielles

(découpage en intervalles -> pris comme modalités)

→ distance/similarité sur tableau disjonctif

- transformation des variables catégorielles en variables numériques

- utilisation de mesures "mixtes »

Principe : $d^2(i, j) = \frac{1}{p} \sum_{k=1}^{p} \delta_j(i, j)$

[0,1] ⇒ Normaliser !!!!

Données mixtes

29

Pour une variable numérique :

$$\delta_k(i, j) = \frac{(x_{ik} - x_{jk})}{(\max - \min)}$$

Distance – Données mixtes

30

Exemple : (Age, Propriétaire résidence principale, montant des mensualités en cours)

x=(30,1,1000), y=(40,0,2200), z=(45,1,4000)

d(x,y)=sqrt($(10/15)^2 + 1^2 + (1200/3000)^2$) = 1.27

d(x,z)= sqrt($(15/15)^2 + 0^2 + (3000/3000)^2$) = 1.41

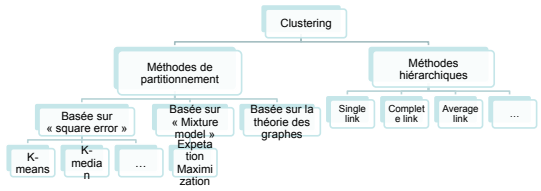
d(y,z)= sqrt($(5/15)^2 + 1^2 + (1800/3000)^2$) = 1.21

plus proche voisin de x = y

Distances normalisées.

Sommation : d(x,y)=d1(x1,y1) + … + dn(xn,yn)

Taxonomie



Algorithmes à partionnement

Construire une partition à k clusters d'une base D de n objets

Les k clusters doivent optimiser le critère choisi

- Global optimal: Considérer toutes les k-partitions
- Heuristic methods: Algorithmes k-means et k-medoids
- k-means (MacQueen'67): Chaque cluster est représenté par son centre
- k-medoids or PAM (Partition around medoids) (Kaufman & Rousseeuw'87): Chaque cluster est représenté par un de ses objets

Algorithme des k-moyennes (K-means)

Entrée : un échantillon de m enregistrements $x_1, \ldots, x_m$

Paramètre : Fixer le nombre de cluster K

1. Choisir k centres initiaux $c_1, \ldots, c_k$
2. Répartir chacun des m enregistrements dans le groupe i dont le centre ci est le plus proche.
3. Si aucun élément ne change de groupe alors arrêt et sortir les groupes
4. Calculer les nouveaux centres : pour tout i, c_i est la moyenne des éléments du groupe i (le barycentre).

Aller en 2.

Illustration (les données)

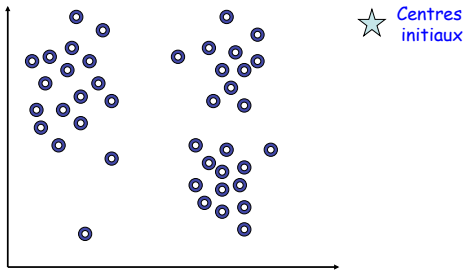


Illustration (Etape 1)

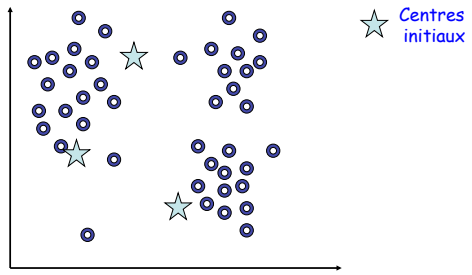


Illustration (Etape 2)

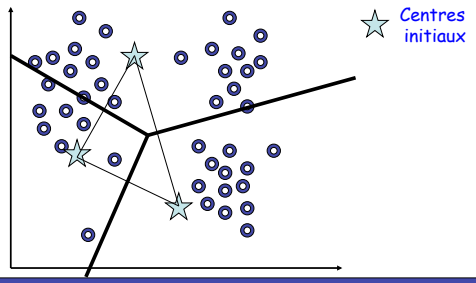


Illustration (Etape 3)

37

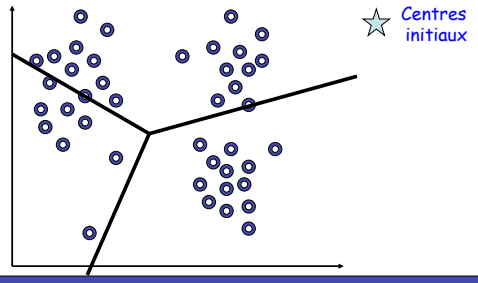


Illustration (Etape 4)

38

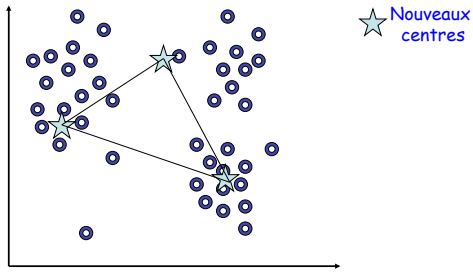


Illustration (Etape 1)

39

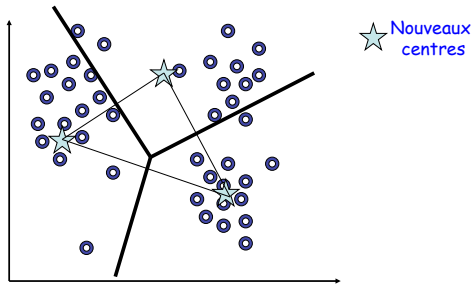
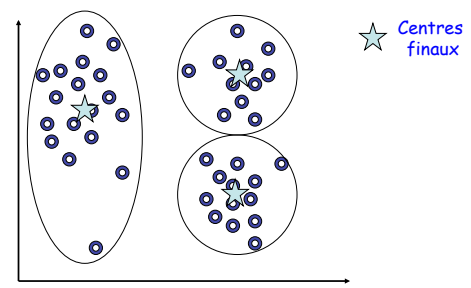


Illustration (3)

40



Algorithme des k-moyennes : Exemple

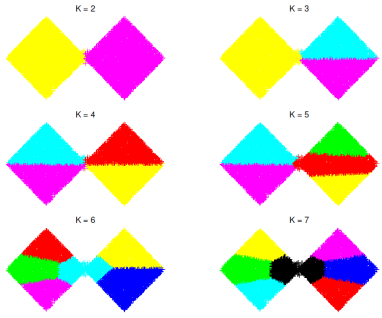
41

- 8 points A, …, H de l 'espace euclidéen 2D. k=2 (2 groupes)
- Tire aléatoirement 2 centres : B et D choisis.

points	Centre D(2,4), B(2,2)	Centre D(2,4), I(27/7,17/7)	Centre J(5/3,10/3), K(24/5,11/5)
A(1,3)	B	D	J
B(2,2)	B	I	J
C(2,3)	B	D	J
D(2,4)	D	D	J
E(4,2)	B	I	K
F(5,2)	B	I	K
G(6,2)	B	I	K
H(7,3)	B	I	K

Exemple

42



Qualité

Mesurer la qualité du clustering

- Compacité des clusters.
- Séparation des clusters.
- Score de la partition.

Compacité

$$wc(C) = \sum_{k=1}^{K} wc(C_k) = \sum_{k=1}^{K} \sum_{x_i \in C_k} d(x_i, c_k)$$

Autre fonction possible :

$$wc(C_k) = \max_i \min_{x_j \in C_k} \{d(x_i, x_j) / x_i \in C_k, x_i \neq x_j\}$$

La plus grande distance minimale entre deux éléments d'un même cluster

Séparation

- Distance entre les centres des clusters :

$$bc = \sum_{1 \leq j < k \leq K} d(r_j, r_k)$$

- Distance entre ensembles :
  - Distance minimale.
  - Distance maximale.
  - Distance moyenne.

Valeur de la partition

Combiner wc (à minimiser) et bc (à maximiser).

Par exemple :

$$\frac{bc}{wc}$$

ou bien :

$$\frac{\alpha bc + \beta wc}{bc + wc}$$

K-moyennes : Avantages

Relativement extensible dans le traitement d'ensembles de taille importante

Relativement efficace : O(t.k.n), où n représente # objets, k # clusters, et t # iterations. Normalement, k, t << n.

Produit généralement un optimum local ; un optimum global peut être obtenu en utilisant d'autres techniques telles que : algorithmes génétiques, …

K-moyennes : Désavantages

Applicable seulement dans le cas où la moyenne des objets est définie

Besoin de spécifier k, le nombre de clusters, a priori

Incapable de traiter les données bruitées (noisy).

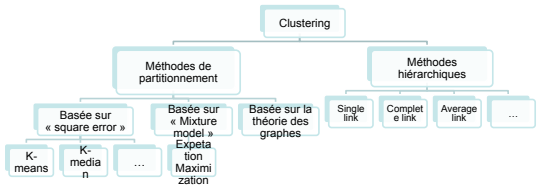
Non adapté pour découvrir des clusters avec structures non-convexes, et des clusters de tailles différentes

Les points isolés sont mal gérés (doivent-ils appartenir obligatoirement à un cluster ?) - probabiliste

K-moyennes : Variantes

- Sélection des centres initiaux
- Calcul des similarités
- Calcul des centres (K-medoids : [Kaufman & Rousseeuw'87])
- GMM : Variantes de K-moyennes basées sur les probabilités
- K-modes : données catégorielles [Huang'98]
- K-prototype : données mixtes (numériques et catégorielles)

Taxonomie



Méthodes hiérarchiques

Une méthode hiérarchique : construit une hiérarchie de clusters, non seulement une partition unique des objets.

Le nombre de clusters k n'est pas exigé comme donnée

Utilise une matrice de distances comme critère de clustering

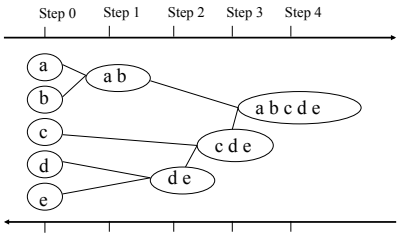
Une condition de terminaison peut être utilisée (ex. Nombre de clusters)

Méthodes hiérarchiques-agglomérative

Entrée : un échantillon de m enregistrements $x_1, \ldots, x_m$

1. On commence avec m clusters (cluster = 1 enregistrement)
2. Grouper les deux clusters les plus « proches ».
3. S'arrêter lorsque tous les enregistrements sont membres d'un seul groupe
4. Aller en 2.

Arbre de clusters : Exemple

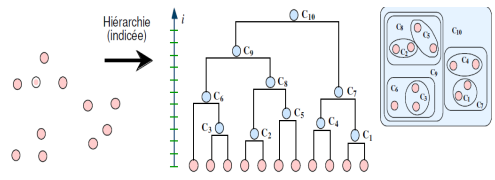


Arbre de clusters - Dendogramme

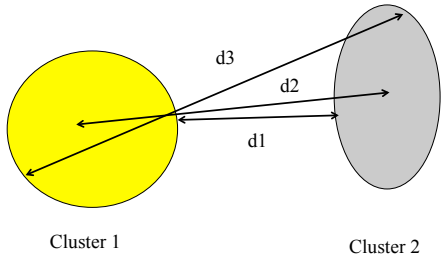
Résultat : Graphe hiérarchique qui peut être coupé à un niveau de dissimilarité pour former une partition.

La hiérarchie de clusters est représentée comme un arbre de clusters, appelé dendrogramme

- Les feuilles de l'arbre représentent les objets
- Les nœuds intermédiaires de l'arbre représentent les clusters



Mais quelle distance ?



Méthode agglomérative

Complexité : n^2

Dépend de la distance entre clusters :

- Distance minimale : clusters allongés.
- Distance maximale : clusters de même volume.
- Distance moyenne.
- Distance entre centres de clusters

Distance entre clusters

Distance entre les centres des clusters (Centroid Method)

- tendance à produire des classes de variance proche

Distance (saut) minimale entre toutes les paires de données des 2 clusters (Single Link Method)

$$d(i,j) = \min_{x \in C_i, y \in C_j} \{ d(x,y) \}$$

- tendance à produire des classes générales (par effet de chaînage)
- sensibilité aux individus bruités.

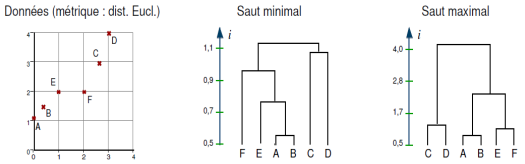
Distance (saut) maximale entre toutes les paires de données des 2 clusters (Complete Link Method)

$$d(i,j) = \max_{x \in C_i, y \in C_j} \{ d(x,y) \}$$

- tendance à produire des classes spécifiques (on ne regroupe que des classes très proches)
- sensibilité aux individus bruités.

Distance (saut) moyenne entre toutes la paires d'enregistrements (Average Linkage) $d(i,j) = avg_{x \in C_i, y \in C_j} \{ d(x,y) \}$

pas les mêmes résultats selon la métrique utilisée ...



Méthodes hiérarchiques : Avantages

Conceptuellement simple

Propriétés théoriques sont bien connues

Quand les clusters sont groupés, la décision est définitive => le nombre d'alternatives différentes à examiner est réduit

Méthodes hiérarchiques : Inconvénients

Groupement de clusters est définitif => décisions erronées sont impossibles à modifier ultérieurement (méthode gloutonne)

Méthodes non extensibles pour des ensembles de données de grandes tailles

61

Clustering : Validation

Critères externes: utilisation de jeux de données dont on connaît le réel regroupement (ex: entropie)

Critères interne : mesure de la qualité sans connaître le réel regroupement (ex: SSE)

Critère relatif

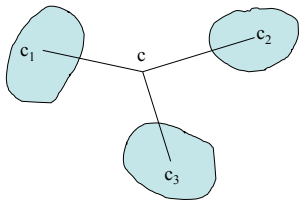
62

Total Sum of Squares (TSS)

$$TSS = \sum dist(x,c)^2$$

$$SSE = \sum_{i=1}^{k} \sum_{x \in C_i} dist(x,c_i)^2$$

$$SSB = \sum_{i=1}^{k} m_i dist(c_i,c)^2$$



c: overall mean

c_i: centroid of each cluster C_i

m_i: number of points in cluster C_i

63

Supervised Cluster Validation: Entropy and Purity

Table 5.9. K-means Clustering Results for LA Document Data Set

Cluster	Entertainment	Financial	Foreign	Metro	National	Sports	Entropy	Purity
1	3	5	40	506	96	27	1.2270	0.7474
2	4	7	280	29	39	2	1.1472	0.7756
3	1	1	1	7	4	671	0.1813	0.9796
4	10	162	3	119	73	2	1.7487	0.4390
5	331	22	5	70	13	23	1.3976	0.7134
6	5	358	12	212	48	13	1.5523	0.5525
Total	354	555	341	943	273	738	1.1450	0.7203

entropy For each cluster, the class distribution of the data is calculated first, i.e., for cluster j we compute p_{ij}, the 'probability' that a member of cluster j belongs to class i as follows: $p_{ij} = m_{ij}/m_j$, where m_j is the number of values in cluster j and m_{ij} is the number of values of class i in cluster j. Then using this class distribution, the entropy of each cluster j is calculated using the standard formula $e_j = \sum_{i=1}^{L} p_{ij} \log_2 p_{ij}$, where the L is the number of classes. The total entropy for a set of clusters is calculated as the sum of the entropies of each cluster weighted by the size of each cluster, i.e., $e = \sum_{i=1}^{K} \frac{m_j}{m} e_j$, where m_j is the size of cluster j, K is the number of clusters, and m is the total number of data points.

purity Using the terminology derived for entropy, the purity of cluster j, is given by $purity_j = \max p_{ij}$ and the overall purity of a clustering by $purity = \sum_{i=1}^{K} \frac{m_i}{m} purity_j$.

64

Pair-counting measures

Measure the number of pairs that are in:

Same class **both** in P and G.

$$a = \frac{1}{2}\sum_{i=1}^{K}\sum_{j=1}^{K'} n_{ij}(n_{ij}-1)$$

Same class in P but different in G.

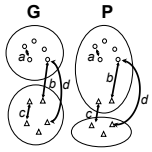
$$b = \frac{1}{2}\left(\sum_{j=1}^{K'} n_{\cdot j}^2 - \sum_{i=1}^{K}\sum_{j=1}^{K'} n_{ij}^2\right)$$

Different classes in P but same in G.

$$c = \frac{1}{2}\left(\sum_{i=1}^{K} n_{i\cdot}^2 - \sum_{i=1}^{K}\sum_{j=1}^{K'} n_{ij}^2\right)$$

Different classes **both** in P and G.

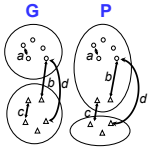
$$d = \frac{1}{2}\left(N^2 + \sum_{i=1}^{K}\sum_{j=1}^{K'} n_{ij}^2 - \left(\sum_{i=1}^{K} n_{i\cdot}^2 + \sum_{j=1}^{K'} n_{\cdot j}^2\right)\right)$$



65

Rand and Adjusted Rand index

[Rand, 1971] [Hubert and Arabie, 1985]



Agreement: a, d
Disagreement: b, c

$$RI(P,G) = \frac{a+d}{a+b+c+d}$$

$$ARI = \frac{RI - E(RI)}{1 - E(RI)}$$

66

External indexes

If true class labels (*ground truth*) are known, the validity of a clustering can be verified by comparing the class labels and clustering labels.

$$\begin{array}{c|c} N & \cdot \\ \hline \cdot & n_{\cdot\cdot} \end{array} = \begin{array}{cccc|c} n_{11} & n_{12} & \dots & n_{1l} & n_{1\cdot} \\ n_{21} & n_{22} & \dots & n_{2l} & n_{2\cdot} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ n_{k1} & n_{k2} & \dots & n_{kl} & n_{k\cdot} \\ \hline n_{\cdot 1} & n_{\cdot 2} & \dots & n_{\cdot l} & n_{\cdot\cdot} \end{array}$$

n_{ij} = number of objects in class i and cluster j

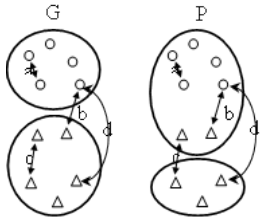
Pointwise measures

$$a = \frac{1}{2}\sum_{i=1}^{K}\sum_{j=1}^{K'} n_{ij}(n_{ij}-1)$$

$$b = \frac{1}{2}\left(\sum_{j=1}^{K'} n_{\cdot j}^2 - \sum_{i=1}^{K}\sum_{j=1}^{K'} n_{ij}^2\right)$$

$$c = \frac{1}{2}\left(\sum_{i=1}^{K} n_{i\cdot}^2 - \sum_{i=1}^{K}\sum_{j=1}^{K'} n_{ij}^2\right)$$

$$d = \frac{1}{2}\left(N^2 + \sum_{i=1}^{K}\sum_{j=1}^{K'} n_{ij}^2 - \left(\sum_{i=1}^{K} n_{i\cdot}^2 + \sum_{j=1}^{K'} n_{\cdot j}^2\right)\right)$$



Rand index

(example)

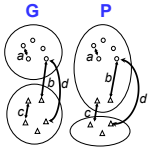
Vectors assigned to:	Same cluster	Different clusters
Same cluster in ground truth	20	24
Different clusters in ground truth	20	72

Rand index = (20+72) / (20+24+20+72) = 92/136 = **0.68**

Adjusted Rand = (to be calculated) = **0.xx**

Pair-counting measures

Agreement: a, d
Disagreement: b, c



$$a = \frac{1}{2}\sum_{i=1}^{K}\sum_{j=1}^{K'} n_{ij}(n_{ij}-1)$$

$$b = \frac{1}{2}\left(\sum_{j=1}^{K'} n_{\cdot j}^2 - \sum_{i=1}^{K}\sum_{j=1}^{K'} n_{ij}^2\right)$$

$$c = \frac{1}{2}\left(\sum_{i=1}^{K} n_{i\cdot}^2 - \sum_{i=1}^{K}\sum_{j=1}^{K'} n_{ij}^2\right)$$

$$d = \frac{1}{2}\left(N^2 + \sum_{i=1}^{K}\sum_{j=1}^{K'} n_{ij}^2 - \left(\sum_{i=1}^{K} n_{i\cdot}^2 + \sum_{j=1}^{K'} n_{\cdot j}^2\right)\right)$$

Rand Index: $RI(P,G) = \frac{a+d}{a+b+c+d}$

Adjusted Rand Index: $ARI = \frac{RI - E(RI)}{1 - E(RI)}$