Indices de qualité en clustering

Stéphane Lallich*, Philippe Lenca**

* Laboratoire ERIC,
Université de Lyon
stephane.lallich@univ-lyon2.fr

** UMR 6285 Lab-STICC,
Telecom Bretagne
Philippe.Lenca@telecom-bretagne.eu

Journée Clustering 2015

20 Octobre 2015

Orange Labs, Issy Les Moulineaux

Plan

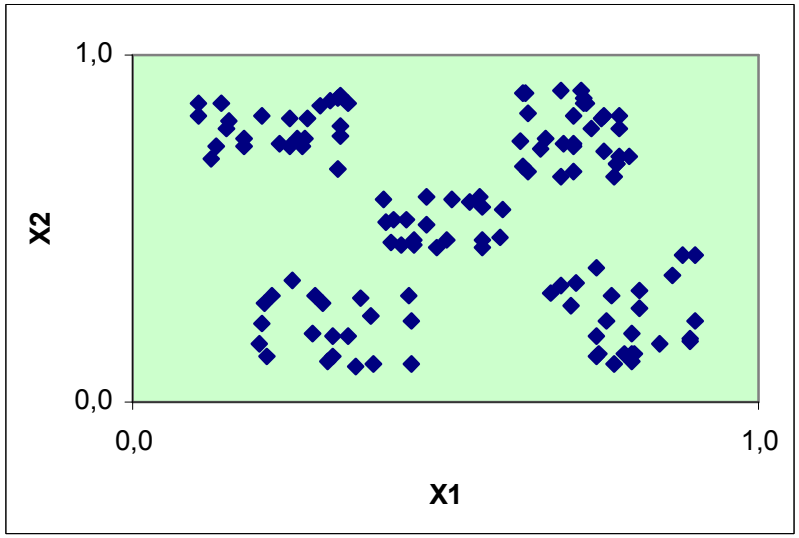
1. Introduction : clustering et qualité
2. Principaux indices internes de qualité d'un clustering
3. Expérimentations pour indices internes
4. Indices internes pour des problèmes spécifiques
5. Approche axiomatique pour le choix d'un indice de qualité
6. Principaux indices externes de qualité d'un clustering
7. Références bibliographiques

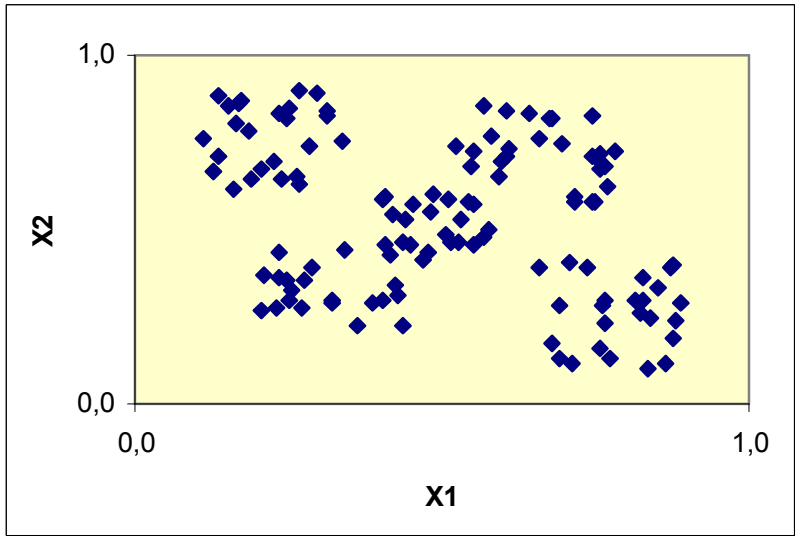
1. Introduction :

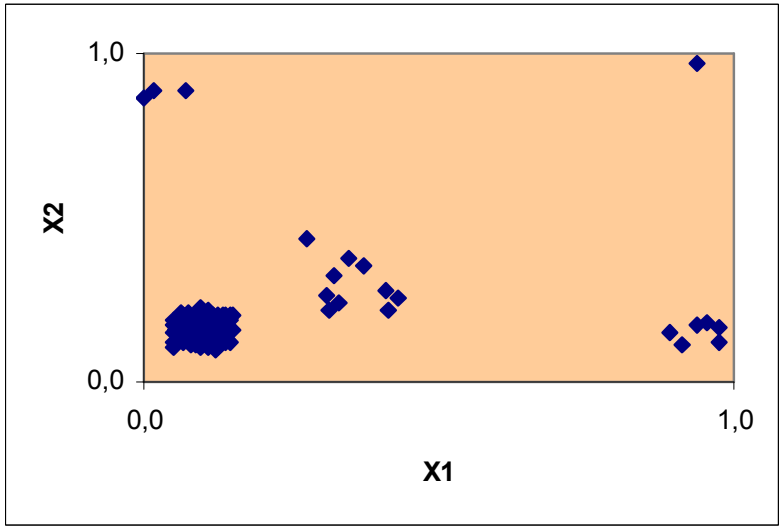
clustering et qualité

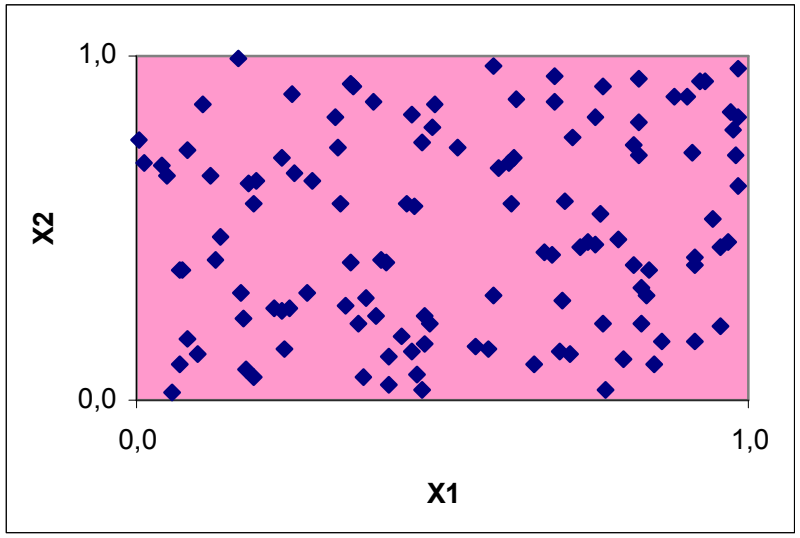
Exemples

4 nuages, n=120 objets et p=2 descripteurs pour chaque nuage



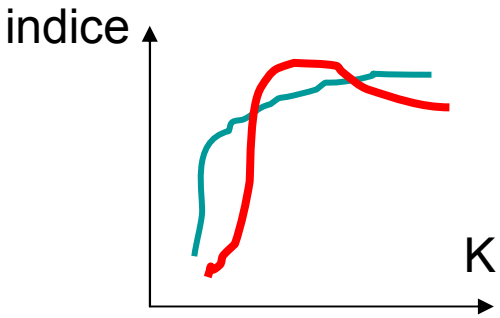






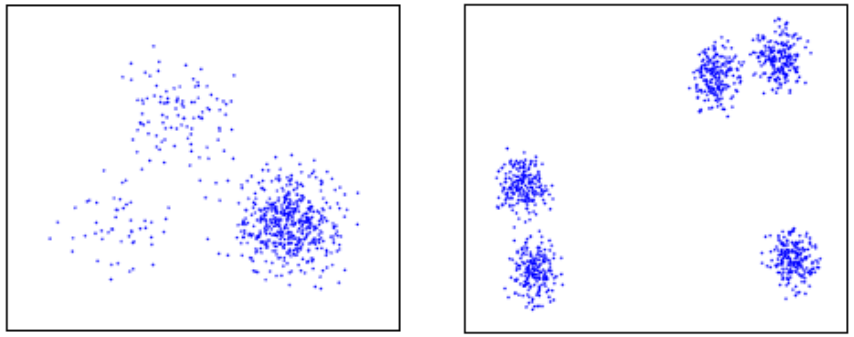
Points-clef pour caractériser un indice de qualité

- détermination du nombre de clusters optimal (variations de l'indice en fonction du nb de clusters)



- compacité des clusters (faible dispersion autour du centre, diamètre)
- séparabilité des clusters (éloignement des centres pris 2 à 2, diamètre de la réunion des deux clusters)
- fidélité aux données de départ
- exemples atypiques, dissymétrie des variables,
- taille des clusters
- complexité du calcul de l'indice
- interprétabilité de l'indice

- réaction face à une densité hétérogène
- gestion des clusters très proches (subclusters)



Extrait de Liu et al. 10

Difficultés

- pas de norme de référence, pas de vérité terrain
- pas de mesure standard comme en supervisé
- pas de distinction *Test Set / Learning Set*

→ **Nécessité** d'indices adaptés au but recherché et aux données !

Différents types d'indice de qualité

Indices internes vs externes

- **Indices internes :** l'évaluation des regroupements obtenus se fait de façon non supervisée à partir des seules données utilisées pour construire les clusters.
→ Nécessité d'indices de qualité internes spécifiques.

- **Indices externes :** l'évaluation du clustering se fait à partir d'informations externes sur les objets, par exemple une étiquette de classe.
→ On est alors ramené aux indices d'une évaluation supervisée. Très commode, mais très contestable !

Indices relatifs

L'évaluation se fait en comparant les résultats de plusieurs clusters ou clusterings, à partir d'indices externes ou internes.

Approche statistique et approche descriptive

On a le choix entre une approche statistique et une approche descriptive (évaluation interne).

Approche statistique à l'aide de tests et de p-values,

- H_0 est le caractère aléatoire pur de la structure. Voir Jain et Dubes pour les différentes formulations de H_0
- la loi de l'indice sous H_0 est svt établie par simulation
- le coût de calcul est très élevé

Approche descriptive. On se limite au calcul de la valeur de l'indice de qualité

L'approche statistique est souvent décevante, car il est difficile d'accepter H_0 et les p-values sont très petites. L'approche descriptive est donc souvent préférée.

2. Principaux indices internes de qualité

Une liste de 27 indices

Index	Rule	Name in R	Date
Ball Hall	max diff	Ball, Hall	1965
Banfeld Raftery	min	Banfeld, Raftery	1993
C index	min	Hubert, Schultz	1976
Calinski Harabasz	max	Calinski, Harabasz	1974
Davies Bouldin	min	Davies, Bouldin	1979
Det Ratio	min diff	Scott, Symons	1971
Dunn	max	Dunn	1974
GDI	max	Bezdek, Pal	1998
Gamma	max	Baket, Hubert	1975
G plus	min	Rohlf	1974
Ksq DetW	max diff	Marriott	1975
Log Det Ratio	min diff	Scott, Symons	1971
Log SS Ratio	min diff	Hartigan	1975
McClain Rao	min	McClain Rao	2001
PBM	max	Bandyopadhyay et al.	2004
Point biserial	max	Point biserial	1981
Ratkowsky Lance	max	Ratkowsky Lance	1978
Ray Turi	min	Ray Turi	1999
Scott Symons	min	Scott Symons	1971
SD	min	SD Scat	2001
S Dbw	min	SD Dis	2001
Silhouette	max	Dbw	2001
Tau	max	Rousseeuw	1987
Trace W	max diff	Edwards, Cavalli-Sforza	1965
Trace WiB	max diff	Friedman, Rubin	1967
Wemmert Gancarski	max	Wemmert, Gancarski	
Xie Beni	min	Xie, Beni	1991

Se reporter aux packages de R :

- **ClusterCrit**

Desgraupes 13

← les 27 indices

- **NbClust**

Charrad et al. 14

30 indices

Inerties *within* et *between*

***I_{tot}*, l'inertie totale du nuage** est la somme des carrés des distances des objets au centre du nuage.

$$I_{tot} = \sum_{i=1}^{n} \| x_i - c \|^2 = \sum_{i=1}^{n} d^2(x_i, c)$$

Décomposition : $I_{tot} = W + B$

W mesure la cohésion des clusters par la somme des carrés des distances de chaque objet à son centre de cluster
→ **le plus petit souhaité**

$$W = \sum_{i=1}^{n} \| x_i - c(x_i) \|^2 = \sum_{i=1}^{n} d^2(x_i, c(x_i))$$

B mesure la séparation des clusters par la somme des carrés des distances entre centres → **le plus grand souhaité**

$$B = \sum_{k=1}^{K} n_k \| c_k - c \|^2 = \sum_{k=1}^{K} n_k d^2(c_k, c)$$

Ratio B/W et indice CH, cf. Calinski et Harabasz 74

Ces mesures intègrent *B* et *W* en un seul indice, pénalisé par le nombre de classes, noté *K*, dans le cas de *CH*

$$\text{ratio } B/W = \frac{B}{W} \qquad CH = \frac{(n-K)B}{(K-1)W}$$

- plus les classes sont compactes et séparables, plus *B* est grand, *W* petit et donc ratio *B/W* et *CH* grand.
- *CH* pénalise d'autant plus le ratio *B/W* que *K* est grand, ce qui facilite la détermination du nombre de classe optimal
- *CH* est à maximiser, compris entre 0 et + ∞, sensible au bruit
- C'est l'indice ayant obtenu les meilleurs résultats dans l'étude de Milligan et Cooper 85.

DB, Indice de Davies-Bouldin, cf. Davies et Bouldin 79

L'indice DB repose sur la notion de similarité de 2 clusters

Similarité de 2 clusters

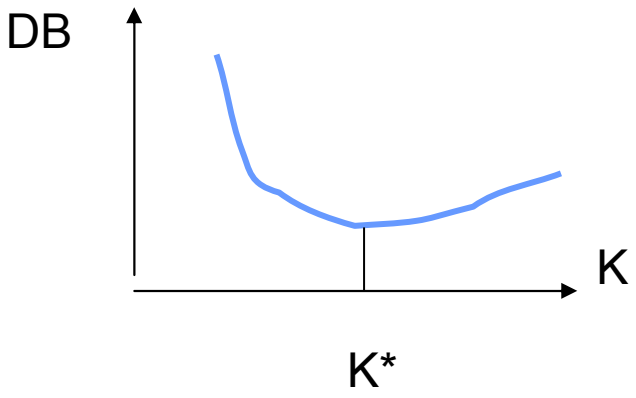
- dispersion d'un cluster k : δ_k
- dissimilarité de 2 clusters : $\Delta_{kk'}$
- similarité des clusters k et k' : $R_{kk'} = \frac{\delta_k + \delta_{k'}}{\Delta_{kk'}}$

Définition de DB

- dispersion du cluster k → δ_k = distance moyenne des objets du cluster k au centre du cluster
- dissimilarité de 2 clusters → $\Delta_{kk'}$ = distance du centre du cluster k au centre du cluster k'
- formule de DB : $DB = \frac{1}{K} \sum_{k=1}^{K} Max_{k' \neq k} \{R_{kk'}\}$

Utilisation de *DB*

- Interprétation : *DB* est la moyenne des similarités entre chaque cluster et le cluster le plus similaire
- la qualité du clustering est d'autant plus grande que *DB* est petit
- *DB* ne dépend pas du nombre de classes



- des variantes de *DB* associées aux différents types de graphes de voisinage (*MST, RNG, GG*) sont proposées par Pal et Biswas 97,
- deux autres variantes sont proposées par Kim et al. 05

DI, GDI, indices de Dunn, cf. Dunn 74, Bezdek et Pal 98

Pour évaluer la **séparabilité des clusters** :

- on évalue la distance de 2 clusters C_k et $C_{k'}$ par la distance de leurs plus proches voisins
- puis on forme d_{min}, la plus petite de ces distances

$$d_{kk'} = Min_{x \in C_k, y \in C_{k'}} \{d(x,y)\}$$

$$d_{\min} = Min_{k' \neq k} \{d_{kk'}\}$$

Afin de prendre en compte la **compacité des clusters** :

- on calcule le diamètre $D(C_k)$, de chaque cluster C_k
- on en déduit D_{max}, le plus grand diamètre de classe.

$$D(C_k) = Max_{x,y \in C_k} \{d(x,y)\}$$

$$D_{\max} = Max_{k=1,2,..K} \{D(C_k)\}$$

Indice de Dunn

$$DI = \frac{d_{\min}}{D_{\max}}$$

Défauts : sensibilité au bruit, complexité de calcul

GDI, index de Dunn généralisé (18 variantes, Bezdek, Pal 98)

$$GDI = \frac{Min_{k \neq k'} \left\{ \delta_{k,k'} \right\}}{Max_k \left\{ \Delta_k \right\}}$$

- $\delta_{kk'}$ distance entre clusters (séparabilité)

6 variantes proposées

- Δ_k mesure la dispersion du cluster k (compacité)

3 variantes proposées

- au total 18 variantes, voir formules dans Desgraupes 13

SIL, indice Silhouette, cf. Rousseeuw, 87

Silhouette d'un objet i

$$SIL(i)_i = \frac{b_i - a_i}{\max(a_i, b_i)}$$

a_i, distance moyenne de l'objet i aux objets de son cluster → **Compacité**

$$a_i = \frac{1}{n(C(x_i)) - 1} \sum_{x_j \in C(x_i)} d(x_i, x_j)$$

b_i, minimum des distances moyenne de l'objet i aux objets de chacune des autres classes
→ **Séparabilité**

$$b_i = \underset{k' \neq k(x_i)}{Min} \left\{ \frac{1}{n_{k'}} \sum_{x_j \in C_{k'}} d(x_i, x_j) \right\}$$

Silhouette d'un cluster C_k ou d'un clustering C
= moyenne des silhouettes concernées, SIL(C_k) ou SIL(C)

SIL est compris entre -1 est 1, à maximiser. Très bons résultats lors des expériences d'Arberlaitz 13.

CI, C-index, cf. Dalrymple-Alford 70

Nombre de distances entre objets 2 à 2, total, intra, inter

$$N_{tot} = \frac{n(n-1)}{2} \qquad N_W = \sum_{k=1}^{K} \frac{n_k(n_k-1)}{2} \qquad N_B = N_{tot} - N_W$$

C-index

S_W : somme des N_W distances intra-cluster
S_{min} : somme des N_W distances les plus petites parmi les N_{tot}
S_{max} : somme des N_W distances les plus grandes parmi les N_{tot}

$$CI = \frac{S_W - S_{\min}}{S_{\max} - S_{\min}}$$

- varie entre 0 (le mieux) et 1 (le pire), à minimiser
- classé 4e/30 par Milligan et Cooper 85
- vérifie les axiomes posés par Ackerman et Ben David 08

Γ, **Indice Gamma,** cf. Hubert, Baker 75, Hubert Arabie 81

Indice Γ_{BH} de Baker et Hubert

Adaptation au clustering du coefficient de corrélation de rangs de Goodman et Kruskal

- S+ : nombre de fois où la distance entre 2 objets d'une paire du type *within* est strictement inférieure à une distance entre 2 objets d'une paire *between* (concordance)
- S- : nombre de discordances (strictement supérieure).
- indice de Baker et Hubert :

$$\Gamma_{BH} = \frac{S^+ - S^-}{S^+ + S^-}$$

- compris entre -1 et 1, à maximiser, top-1 de Milligan 81

Indice Γ_{HA} de Hubert et Arabie

Corrélation de Pearson sur les paires d'objets entre distance et indicatrice de *between*, et non pas corrélation de rang.

RM, Relative margin, cf. Ackerman Ben David 08

Marge relative d'un objet

Pour chaque objet x_i, on calcule $d(x_i,c_1(x_i))$ et $d(x_i,c_2(x_i))$ les distances de x_i à ses 1er et 2e plus proches centre et l'on forme le quotient des 2 distances

$$RM_i = \frac{d\left(x_i, c_1(x_i)\right)}{d\left(x_i, c_2(x_i)\right)}$$

Marge relative

On calcule la moyenne des marges relatives sur les objets

$$RM = \frac{1}{n}\sum_{i=1}^{n} \frac{d\left(x_i, c_1(x_i)\right)}{d\left(x_i, c_2(x_i)\right)}$$

- varie entre 0 (le mieux) et 1 (le pire), à minimiser
- vérifie les axiomes posés par Ackerman et Ben David 08

Indice MB, cf. Maulik, Bandyopadhyay 02

$$MB = \left(\frac{1}{K} \frac{E_1}{E_K} D_K \right)^{\theta}$$

- E_K est la somme des distances séparant chaque objet de son centre de classe → **compacité**

$$E_K = \sum_{k=1}^{K} \sum_{i=1}^{n} u_{ik} d(x_i, c_{k'})$$

E_1 joue le rôle d'un facteur de normalisation, u_{ik} est le t.g. de la matrice de partition U(n,K), θ un paramètre de contrôle.

- D_K est la plus grande des distances entre 2 centres de cluster → **séparabilité**

$$D_K = Max_{k,k'} \{ d(c_k, c_{k'}) \}$$

Quand *K* ↑, *1/K* ↓, E_K ↑ et D_K ↑. *BM* est à maximiser.

3. Expérimentations pour indices internes

Expérimentations

Données synthétiques

→ données générées suivant un nombre de clusters fixé

☺ on contrôle le nombre de clusters, voire leur forme, le bruit…

☹ structures de covariances peu subtiles, loin du réel

Données réelles

→ Jeu de données réelles avec une variable de classe

☺ on se ramène sous le réverbère de l'apprent. supervisé

☹ ici au moins, on a de la lumière …

Données obtenues par resampling

→données réelles, voire synthétiques, rééchantillonnées par bootstrap ou autre …

☺ étude de stabilité

☹ renseigne plus sur la stabilité que sur la qualité

Expérimentations : un tableau de synthèse, cf. Liu et al. 10

Nom	Notation	Mono	Noise	Dens	Sub-C	Skew	Opt
Root-mean-square std dev	RMSSTD	**x**	**---**	**---**	**---**	**---**	Elbow
R-squared	RS	**x**	**---**	**---**	**---**	**---**	Elbow
Modified Hubert Γ	G	**x**	**---**	**---**	**---**	**---**	Elbow
Calinski-Harabasz index	CH		**x**				Max
I index	II			**x**			Max
Dunn's indices	DI		**x**		**x**		Max
Silhouette index	SIL				**x**		Max
Davies-Bouldin index	DB				**x**		Min
SD validity	SD				**x**		Min
S_Dbw validity index	**SDbw**						**Min**
Xie-Beni index	XB				**x**		Min

Expérimentations : autres exemples

Milligan et Cooper 85

30 indices, CAH, 108 jeux synthétiques, classes non empiétantes, en faisant varier le nb de classes, le nb de variables et la taille de clusters

Le top-5 : CH, indice Je(2)/J(1) de Duda, C-Index, Gamma de Baker et Hubert, Beale (analogue à un F-ratio)

Arbelaitz et al. 13

30 indices, 720 jeux de données synthétiques, 20 jeux de données réelles, 3 algorithmes (k-means, Ward and Average-linkage).

Les performances diminuent en cas de chevauchement des classes et de bruit

Le top-6 : Silhouette, Davies-Bouldin, Calinski-Harabasz, generalized Dunn, COP and SDbw|

Levine 01

Resampling

Kim et al. 05

7 jeux synthétique (2 dim), 4 jeux réels, met en avant des améliorations de DB et XB

4. Indices internes pour des problèmes spécifiques

Indices de qualité pour des problèmes spécifiques

- fuzzy clustering : cf. Wang, Zhang 07
- clustering fondé sur les graphes : cf. van Laarhoven et Marchiori 14
- clustering spatio-temporel : cf. Rizoiu et al. à paraître
- clustering semi-supervisé
- clustering avec contrainte de contiguïté
- co-clustering
- clustering à base de règles
- comités de regroupeurs

Fuzzy clustering : Indices fondés sur les d° d'appartenance

Question : quel est le degré de flou d'un fuzzy clustering ?

Notations : m_{ij} indique le d° d'appartenance de l'objet i au cluster j

Fuziness Performance Index, FPI

Coefficient de partition de Bezdek, F

$$F = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{k} m_{ij}^2$$ varie entre 1 (hard) et 1/k (unif)

Fuzziness Performance Index de Roubens, FPI

$$FPI = 1 - \frac{kF - 1}{k - 1}$$ varie entre 0 (hard) et k (uniforme)

Normalized Classification Entropy, NCE

Aussi appelée Modified Partition entropy, MPE

Entropie moyenne, H (entre 0 et log *k*)

$$H = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k} m_{ij}\log m_{ij}$$

Entropie normalisée, NCE (entre 0 et 1)

$$NCE = \frac{H}{\log k}$$

Fuzzy-clustering : indice XB, cf. Xie et Beni 91

- la compacité des clusters est évaluée par *W*, l'inertie *within*, (cf. indice CH) en tenant compte de m_{ik}, le d° d'appartenance de l'objet i au cluster k
- la séparabilité des clusters est évaluée comme dans l'indice de Dunn, par *dmin,* la plus petite des distances entre 2 clusters au sens des plus proches voisins

$$W = \sum_{i=1}^{n} \sum_{k=1}^{K} m_{ik}^2 d^2(x_i, c_k) \qquad d_{\min} = Min_{k' \neq k} \left\{ d_{kk'} \right\}$$

- on obtient ***XB*** en divisant *W* par nd^2_{min} : $XB = \dfrac{W}{nd_{\min}^2}$
- *XB* est à minimiser

5. Approche axiomatique pour le choix d'un critère de qualité

Approche axiomatique pour le choix d'un regroupeur

Formalisation

- **X**, un ensemble de **n** objets
- **d** : X × X → R, une dissimilarité sur X
- $\mathcal{D}$: ensemble des dissimilarités sur X
- **f** : (X, $\mathcal{D}$) → $\mathcal{C}$, un regroupeur sur X, f(X, d) = C
- **C** = {C_1, C_2, … , C_k} une partition en **k** clusters issue de f
- $\mathcal{C}$: ensemble des partitions de X en k clusters
- $\forall$ (x, y) $\in$ XxX, **x C-eq y** ssi x et y sont dans le même cluster
- **m** : $\mathcal{C}$ → R+, mesure de qualité d'un regroupement

Axiomes pour un regroupeur, Kleinberg 02

1. Invariance d'échelle : f(d) = f(λd)
où λd défini par (λd)(x, y) = $\lambda \times$d(x, y), λ > 0, $\forall$ (x, y) $\in$ XxX

2. Cohérence : le résultat du regroupement ne doit pas changer si les distances intra-clusters sont diminuées et les distances inter-clusters augmentées

- Variante C-cohérente de d : $\forall$ C sur (X, d), une dissimilarité d' est une variante C-cohérente de d si d'(x, y) < d(x, y) pour tout x C-eq y et si d'(x, y) > d(x, y) pour tout x non C-eq y.
- Alors f est cohérente si f(d) = f(d'), quelle que soit d', variante C-cohérente de d.

3. Richesse : le regroupeur f est riche si n'importe quelle partition de X peut être obtenue en modifiant la dissimilarité d : $\forall C \in \mathcal{C}$, $\exists$ d | f(d) = C.

Théorème d'impossibilité de Kleinberg : aucune fonction de regroupement ne satisfait à la fois les 3 axiomes proposés

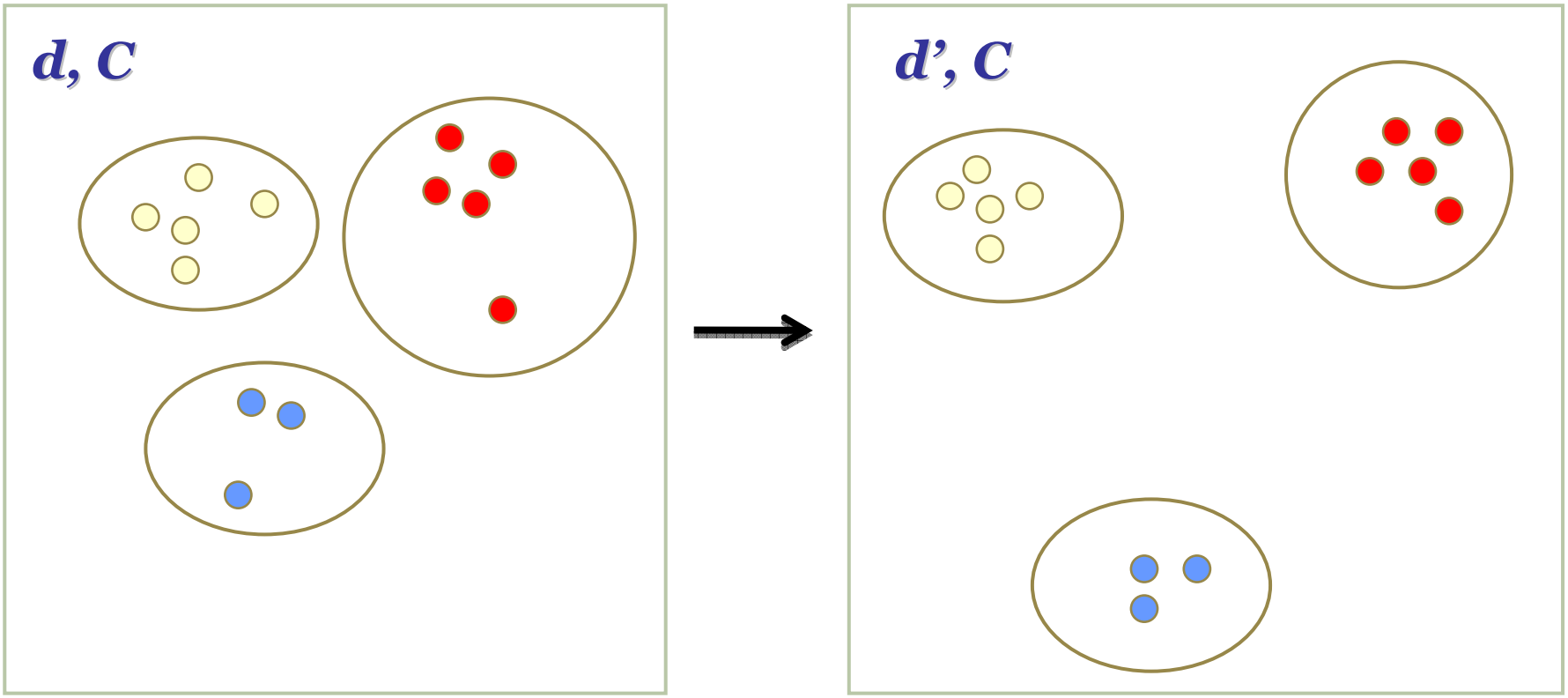


Illustration de la cohérence, extrait d'Ackerman et Ben-David

Axiomes pour un indice de qualité (Ack, Ben 08)

1. Invariance d'échelle. Un changement d'échelle sur les distances ne modifie pas la qualité du clustering

$\forall C \in \mathcal{C}$, C = f(X, d), $\forall\ \lambda > 0$, m(C, X, λd) = m(C, X, d).

2. Cohérence. L'indice m augmente si l'on remplace d par d', une dissimilarité C-cohérente avec d.

$\forall C \in \mathcal{C}$, $\forall$d' $\in \mathcal{D}$, d' C-coh avec d, m(C, X, d') $\geq$ m(C, X, d)

3. Richesse $\forall C \in \mathcal{C}^*$, $\exists$ d $\in$ D | C = Argmax {m(C, X, d)}

Pour chaque partition non triviale C de X, il existe une dissimilarité d qui produit C.

Théorème d'existence : L'invariance d'échelle, la cohérence et la richesse constituent un ensemble cohérent d'exigences pour une mesure de qualité

Exemples : Relative margins, Additive margins, Weakest link

4. Un axiome additionnel, l'invariance permutationnelle

Pour donner à leur ensemble d'axiomes plus de pertinence **(correct & complet),** Ackerman et Ben David ajoutent l'axiome d'invariance par isomorphisme préservant les distances

Partitions isomorphes. Deux partitions C et C' sur le même domaine (X, d) sont isomorphes ssi il existe $\varphi : X \rightarrow X$, un isomorphisme sur X qui préserve les distances, tel que
x C-eq y $\Leftrightarrow$ φ(x) C'-eq φ(y)

Invariance par isomorphisme préservant les distances.
L'indice m n'est pas affecté par un isomorphisme sur X qui préserve les distances
$\forall$C, C' $\in$ C , C et C' isomorphes, m(C, X, d) = m(C', X, d)

L'indice ne change pas en cas de permutation des étiquettes des objets.

Indices vérifiant les axiomes d'Ackerman et Ben-David

- **C-index** (Dalrymple-Alford, 70)
- **Gamma** (Baker et Hubert, 75)
- **Adjusted ratio of clustering** (Roenker et al., 71)
- **D-index** (Dalrymple-Alford, 70)
- **Modified ratio of repetition** (Bower, Lesgold, Tieman, 69)
- **Dunn's index** (Dunn, 73)
- **Variations of Dunn's index** (Bezdek, Pal, 98)
- **Strict separation** (fondé sur Balacan, Blum, Vempala, 08)

6. Principaux indices externes de qualité

Accuracy, Rappel, Précision, F-mesure et autres

Procédure

Si on dispose d'objets étiquetés, il est possible d'appliquer les procédures d'évaluation habituelles en supervisé.

- clustering sur LS, sans tenir compte de l'étiquette
- prédiction à partir de l'étiquette majoritaire dans le cluster ou de règles adaptées au déséquilibre des classes
- calcul sur TS des indices de qualité habituels à partir de la matrice de confusion booléenne qui croise prédit et réel.
- on préfère le rappel et la précision à la seule accuracy qui est inadaptée au cas de classes déséquilibrées
- la F-mesure permet des comparaisons, mais elle n'apporte pas de connaissances.
- en médecine, on utilise plutôt la sensibilité et la spécificité, ainsi que la valeur ajoutée.

Matrice de confusion, Accuracy, Rappel, Precision, F-mesure

Accuracy :
p = *(TP + TN)/n = (a+d)/n*

Precision :
p = *TP/(TP + FP) = a/(a+c)*

Recall :
r = *TP/(TP + FN) = a/(a+b)*

Predict. Actual	Class 1	Class 2	**total**
Class 1	a (TP)	b (FN)	a+b
Class 2	c (FP)	d (TN)	c+d
total	a+c	b+d	n

F1-mesure : compromis entre *p* and *r*, défini comme la moyenne harmonique de *p* et *r*. Identique à l'indice d'association de *Czekanowski* (ou indice de Dice) entre étiquette prédite et étiquette réelle.

F1-measure = 2TP/(2TP+FN+FP) = 2a/(2a+b+c)

Pureté et information mutuelle normalisée

Pureté : c'est l'accuracy de la prédiction de la classe par le clustering.

Problème : la pureté dépend du nombre de classes !

Pour pallier ce problème, on préfère **l'information mutuelle normalisée, NMI,** qui est comprise entre 0 et 1 et qui ne dépend pas du nombre de classe

$$NMI(C,E) = \frac{2 \times I(C,E)}{H(C) + H(E)}$$

où H(E) est l'entropie de la distribution des étiquettes, H(C) est l'entropie de la distribution des clusters, I (E, C) est l'information mutuelle de E et C.

Quelques autres indices externes

Voir définition dans Desgraupes 13

- Czekanowski-Dice
- Folkes-Mallow
- Hubert
- Jaccard
- Kulczynski
- McNemar
- Phi
- Rand et Rand corrigé
- Rogers-Tanimoto
- Russel-Rao
- Sokal-Sneath

7. Références bibliographiques

Sur le clustering en général

Jain A.K and Dubes R. (1988), Algorithms for Clustering Data, Prentice Hall, Englewood Cliffs, New Jersey ; téléchargeable

Kumar V. (2006), Cluster analysis basic concepts and algorithms. In Introduction to Data Mining, Addison-Wesley, USA, chap. 8, pp. 487-567.

Lanzi, P. L, Clustering, Data Mining and Text Mining (UIC 583 @ Politecnico di Milano), 4 diaporamas téléchargeables

Lebarbier E. et T. Mary-Huard, Classification non supervisée, Cours AgroPArisTech, http://www.agroparistech.fr/Supports-de-cours,1177.html

Theodoridis S., Koutroumbas K. (2003), Pattern recognition, 2nd ed., accessible par le web …

Sur les indices de qualité

- Arbelaitz O., Gurrutxaga I., Muguerza J., Pérez J. M., Perona I. (2013), An extensive comparative study of cluster validity indices, Pattern Recognition, (46)1, 243–256
- Baker F. B, Hubert L. J. (1975). Measuring the Power of Hierarchical Cluster Analysis. Journal of the American Statistical Association, 70(349), 31-38.
- Bezdek J. C., N. R. Pal, Some new indexes of cluster validity, IEEE Transactions on Systems, Man and Cybernetics, Part B 28 (1998) 301-315
- Dalrymple-Alford, E. C. (1970). The measurement of clustering in free recall, Psychological Bulletin, 75, 32-34

Fukuyama Y., Sugeno M. (1989), A new method of choosing the number of clusters for the fuzzy c-means method, Proc. 5th Fuzzy Systems Symp., 247–250.

- Halkidi M., Batistakis Y., Vazirgiannis M. (2002). Cluster Validity Methods: Part I. SIGMOD Record 31(2), pp. 40-45
- Halkidi M., Batistakis Y., Vazirgiannis M. (2002). Clustering Validity Checking Methods: Part II. SIGMOD Record 31(3), pp. 19-27
- Halkidi M., Gunopulos D., Vazirgiannis M., Kumar N., Domeniconi C. (2008). A clustering framework based on subjective and objective validity criteria. TKDD 1(4)
- Halkidi M., Varzigianis M. (2002), An Introduction to Quality Assessment in DataMining –Tutorial, PKDD 02. http://www.unipi.gr/faculty/mhalk/Publ_Maria.htm

- Halkidi M., Vazirgiannis M. (2005). Quality Assessment Approaches in Data Mining. The Data Mining and Knowledge Discovery Handbook, pp. 661-696
- Kim M., R. S. Ramakrishna, (2005), New indices for cluster validity assessment, Pattern Recogn. Lett., vol. 26, no. 15, pp. 2353–2363
- Levine E., Domany E. (2001) Resampling Method for Unsupervised Estimation of Cluster Validity, Letters
- Lévine E., Domany E. (2001), Resampling method for unsupervised estimation of cluster validity. In Neural Comput.13(11):2573-93
- Liu Y., Li Z., Xiong H., Gao X., Wu J. (2010). Understanding of Internal Clustering Validation Measures. In 2010 IEEE International Conference on Data Mining, pp. 911, 916
- Maulik, Bandyopadhyay (2002), Performance evaluation and some clustering algorithms and validity index, IEEE PAMI, vol. 24, pp. 1650-1654
- Milligan G. W.. (1981), A Monte Carlo study of thirty internal criterion measures for cluster analysis. Psychometrika, 46, no. 2:187-199
- Milligan G. W., Cooper M. C. (1985), An examination of procedures for determining the number of clusters in a data set, Psychometrika 50 (2), 159-179
- Milligan G. W., Cooper M. C. (1986), A Study of the Comparability of External Criteria for Hierarchical Cluster Analysis, Multivariate Behavioral Research, Vol. 21, No. 4, pp. 441-458

- Rizoiu M. A., Velcin J., Bonnevay S., Lallich S. (à paraître), ClusPath: A Temporal-driven Clustering to Infer Typical Evolution Paths, Data Mining and Knowledge Discovery Journal
- Tibshirani, R., G. Walther, and T. Hastie. "Estimating the number of clusters in a data set via the gap statistic." *Journal of the Royal Statistical Society: Series B*. Vol. 63, Part 2, 2001, pp. 411–423
- van Laarhoven T., Marchiori E. (2014) Axioms for Graph Clustering Quality Functions. In Journal of Machine Learning Research 15, pp. 193-215.
- Wang W., Zhang Y. (2007), On fuzzy cluster validity indices, Fuzzy Sets and Systems, Vol. 158, Issue 19, 2007, Pages 2095–2117

Implémentation

- R package NbClust, Charrad M., Ghazzali N., Boiteau V., Niknafs A. (2015), Package 'NbClust', Determining the Best Number of Clusters in a Data Set, https://sites.google.com/site/malikacharrad/research/nbclust-package
- R packageClusterCrit, Desgraupes B. (2013), Clustering Indices, University Paris Ouest, Lab Modal'X, cran.r-project.org/web/
- R package ClusterStability, Lord, Lapointe, Makarenkov (2015), https://cran.r-project.org/web/packages/ClusterStability/ClusterStability.pdf

- Mathlab, http://fr.mathworks.com/help/stats/cluster-evaluation.html
- R package clv, Nieweglowski L (2014), clv: Cluster Validation Techniques. R package version 0.3-2.1, http://CRAN.R-project.org/package=clv.

Sur l'approche axiomatique

- Ackerman M., Ben-David S., Loker D. (2010). Towards Property-Based Classification of Clustering Paradigms. Neural Information Processing Systems Conference (NIPS 2010).
- Ackerman M., Ben-David S., Loker D. (2010). Characterization of Linkage-Based Clustering. 23rd International Conference on Learning Theory, COLT'10.
- Ackerman M., Ben-David S. (2009). Clusterability: A Theoretical Study. Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics, AISTATS 2009, JMLR: W&CP 5, pp. 1-8.
- Ackerman M., Ben-David S. (2008).. Measures of Clustering Quality: A Working Set of Axioms for Clustering. Neural Inf. Proc. Syst. Conference (NIPS 2008)
- Kleinberg, J. (2002). An impossibility theorem for clustering. Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press.
- Meila M. (2005), Comparing Clusterings: An Axiomatic View, In ICML '05: Proceedings of the 22nd international conference on Machine Learning, pp. 577–584, New York, NY, USA, ACM Press.