Classification connexe

Catherine Aaron

SAMOS-MATISSE
Université Paris 1
75013 Paris
catherine_aaron@hotmail .com

RÉSUMÉ. On s'intéresse ici à la construction d'une méthode de classification non supervisée sous la seule hypothèse de connexité des classes. Cette hypothèse est l'une des plus générales qu'on puisse faire sur la forme des classes. Après avoir défini une notion de connexité adaptée a des espaces discrets on montrera que la classification hiérarchique par la distance minimum mène à l'obtention de classes connexes. On définira alors une distance intra classe rendant compte de la connexité afin de mettre au point une méthode de choix du nombre de classes.

MOTS-CLÉS : classification, non-supervisée, connexité

1. Algorithme de classification

On dispose d'un ensemble de N points de $\mathbb{R}^P$: $E = \{x_1, ..., x_N\}$ que l'on souhaite segmenter en classes connexes du point de vue d'une distance d .Etant donnée la nature discrète du problème la notion de connexité, capacité à lier deux points d'un ensemble par un chemin continu de point de l'ensemble, doit être adaptée aux espaces discrets.

1.1. N*otion de connexité et espaces discrets*

On définit ici une notion de connexité par seuil comme il suit : E est $\delta - connexe$ si et seulement si on peut lier tous ses couples de points par un chemin constitué de points de E deux à deux distants d'au plus δ .

On montre alors que pour tout δ il existe une unique partition $\delta - connexe$ minimale (au sens ou le nombre de classe est minimal, soit encore que tout regroupement de classes n'est pas $\delta - connexe$). On notera alors $p(\delta)$ le nombre de classe associé à une telle segmentation. La fonction p qui, à δ , associé $p(\delta)$ est alors une fonction en escalier, décroissante, à valeur dans $\{1, ..., N\}$ On notera :

- $\delta_{\min}(k) = borne \inf\{\delta / p(\delta) = k\}$: plus petit seuil pour un nombre de classe donné.
- $\delta_{\max}(k) = borne \sup\{\delta / p(\delta) = k\}$: plus grand seuil pour un nombre de classe donné.

Avec $\delta_{\min}(k) = \delta_{\max}(k+1)$

1.2. C*lassification hiérarchique associée*

On montre que la classification hiérarchique par la distance minimum[1] entre deux ensembles va mener à l'obtention de toutes les classifications $\delta - connexes$ minimales possibles. En effet pour passer d'une classification minimale en $k+1$ classe à une classification minimale en k classes, on montre que la seule possibilité est le regroupement des deux classes les plus proches au sens de la distance minimum. Ce qui correspond à l'algorithme de classification de classification hiérarchique par la distance min.

[1] Les classifications hiérarchiques ascendantes et descendantes sont équivalentes dans le cas du choix de la distance minimum

2. Distance intra-classe

Une fois l'ensemble des segmentations connexes maximales effectuées on souhaite obtenir un indicateur permettant de choisir les segmentations les plus significatives. Pour cela on se base sur les méthodes classiques de minimisations de la distance intra classe, mais ici les notions de distances intra-classes moyennes de type euclidienne ou, de manière similaire les minimisations de variances intra classes, ne sont pas pertinentes.

2.1. *Distance entre deux points*

Pour tenir compte de la notion de connexité, la distance entre deux points doit faire intervenir tous les autres points de l'ensemble. L'exemple ci-contre illustre ce propos : dans les deux cas les points A et B sont à la même distance (euclidienne) mais, du point de vue de la connexité, les situations sont très différentes.

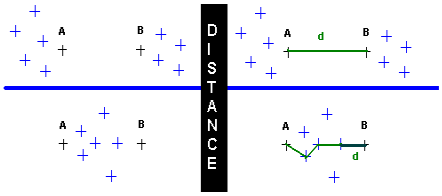

Figure 1. *Distance entre deux points compatible avec la notion de connexité*

Pour résoudre ce problème on choisira, pour distance entre deux points dans un ensemble E le plus petit des sauts maximums effectués lorsqu'on lie les points par un chemin de points de E :

$$d_c^E(x_i,x_j)=\min_{\pi\in chem(x_i,x_j)}\left(\max_{1\le i<N}(d(x_{\pi(i)},x_{\pi(i+1)})\right)$$

avec $chem(x_i,x_j)=\{\pi,\ \text{application de } \{1,....\text{N}\} \text{ dans } \{1,....\text{N}\} \text{ avec : } \pi(1)=i,\pi(N)=j\}$

En notant $Cl_k(i)$ la classe de l'élément x_i de E obtenue lors d'une partition minimale en k classes connexes on a :

$d_c^E=\delta_{\max}(\arg\min\{k / Cl_k(i)=Cl_k(j)\})$ soit le plus petit seuil de connexité pour lequel les points x_i et x_j sont agrégés.

2.2. *Distance intra classe*

Une fois définie d_c on obtient aisément une distance intra-classe pour une segmentation en k classes par : $D_{\text{intra}}(k)=\dfrac{1}{N_k}\sum_i\sum_{\substack{j\neq i\\ x_j\in C_k(i)}} d_c(x_i,x_j)$ avec $N_k=\sum_i\sum_{\substack{j\neq i\\ x_j\in C_k(i)}}1$

On a aussi : $D_{\text{intra}}(k)=\dfrac{\sum_{i=k}^{N-1}\hat{N}_i\delta_{\max}(k)}{\sum_{i=k}^{N-1}\hat{N}_i}$

Avec $\widehat{N}_k=\sum_{i=1}^{N}\sum_{j=1}^{N}1_{\{d_c^E(x_i,x_j)=\delta_{\max}(k)\}}$ soit le nombre de couples distants de $\delta_{\max}(k)$

Une telle écriture de la distance intra classe permet le calcul de toutes les distances intra classes des segmentations successives à l'issue de la classification hiérarchique sans augmentation significative du temps de calcul.

Pour déterminer le nombre de classes optimum on se propose de rechercher la segmentation correspondant à la plus grande rupture de distance intra classe

3. Résultats

3.1. *Classes connexes et convexes*

Dans les exemples suivants : séparation de gaussiennes, base de Ruspini ou Iris de Fischer, les classes sont à la fois connexes et convexes et peuvent, en conséquence, être séparée par d'autres méthodes de classification (voir figure 2)

3.2. *Classes connexes et non convexes*

Dans cet exemple : reconnaissance de classes formées par des cercles imbriqués et bruités la seule caractéristique des classes est la connexité, le regroupement autour de barycentres, comme dans les $K-means$ sera inopérante. On voit ici que la classification hiérarchique permet de retrouver les classes initiales tant que le bruitage n'est pas trop important et que, de plus, le critère du maximum de saut de distance intra est un bon indicateur du nombre de classes (voir figure 3)

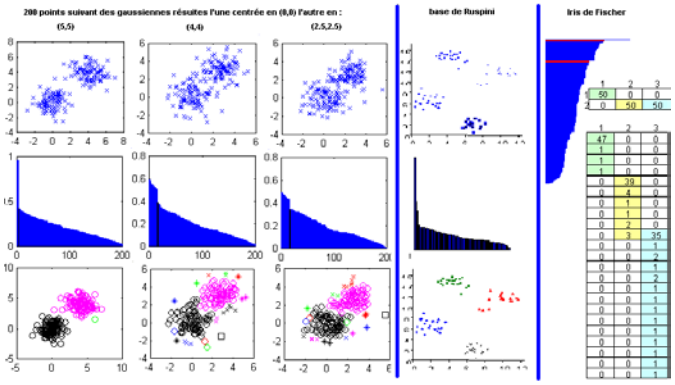

Figure 2.

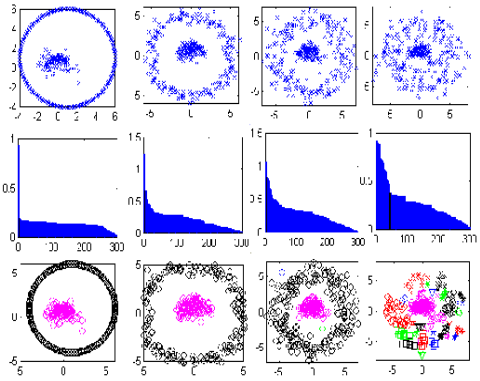

Figure 3.

Exemples de résultats sur des classes connexes, pour chaque exemple excepté les Iris de Fisher on lit de haut en bas : représentation de la base initiale, diagramme des distances intra et résultats de la classification. Dans le cas des Iris de Fischer, la dimension de l'espace étant 4 on ne représente que le diagramme des distances intra et le croisement entre les classifications retenues et les « vraies classes »

4. Construction d'un test de significativité sur la rupture de distance intra classe dans le cas gaussien

Du fait de la complexité de l'expression de la distance intra classe, le taux de significativité du saut maximum de distance intra classe : $saut = \max_{k\in\{1\ldots N/2\}} \frac{D_{\text{intra}}(k-1) - D_{\text{intra}}(k)}{D_{\text{intra}}(k-1)}$ [2] va être estimé par simulation. Dans le cas présent (test d'existence d'une unique classe de forme gaussienne) on a effectué, pour des tailles d'échantillon (N) variant de 5 à 100 individus et pour des dimensions (P) variant entre 1 et 10, 500 tirages sur lesquels on a estimé la valeur du saut.

On remarque empiriquement que :

- Les moyennes et écarts types des sauts diminuent en fonction de N et P. Des estimations de la moyenne et de l'écart type par des fonctions du type :

$$x(N,P) = \exp[-(a\ln(N) + b(\ln(P) + c\ln(N)\ln(P) + d)]$$

donnent de très bons résultats (R2 de 99,8% pour la moyenne et de 98% pour l'écart type, T de Student supérieurs à 9 pour la moyenne, à 4,6 pour l'écart type). Voire tableau 1 pour les résultats numériques.

[2] On prend le maximum sur les premières classifications (de 1 à N/2) pour éviter les problèmes de bords qui apparaissent lorsqu'on scinde la base en un nombre trop élevé de classes.

- La distribution des sauts (centrés normés) semble vite converger en N et en P avec. On estime alors ces différents quantiles d'ordre α (voir tableau 2)

	a	b	c	d
Moyenne	*0.21*	*0.26*	*0.24*	*0.42*
Ecart type	*0.18*	*0.31*	*0.38*	*1.1*

Tableau 1. *Estimation de la moyenne et de l'écart type du saut*

α	90%	95%	97%	99%
Q_α	*1.5*	*2*	*2.4*	*3.2*

Tableau 2. *Quantiles d'ordre alpha de la répartition du saut centré normé en fonction de N et P*

On construit ainsi un test de rejet de l'hypothèse d'existence d'une unique classe de forme gaussienne à $\alpha\%$ si la valeur centrée réduite du saut dépasse le quantile d'ordre α

5. limites de la méthode

Dans le cas où les « vraies » classes C_i vérifieraient des conditions de non séparabilité conjointe, c'est à dire que pour isoler une classe on est obligé d'en scinder une autre, soit encore que :

$$\exists i \neq j \,/\, \exists k \,/ \quad \min_{x_i \in C_i, x_j \in C_j} (d(x_i, x_j)) < \max_{x_{k_1} \in C_k} (\min_{x_{k_2} \in C_k} (d(x_{k_1}, x_{k_2}))$$

Alors la classification hiérarchique ne pourra, en aucuns cas, retrouver conjointement les bonnes. Dans les meilleurs cas (par exemple les exemples des trois premières séparations de gaussiennes, les iris de Fischer ou les premiers cercles concentriques) on isolera les points les plus éloignés de leur classe pour obtenir, au final, une classification satisfaisante. Dans le pire des cas, si : $\exists i \neq j \,/\, \exists k \,/ \quad \min_{x_i \in C_i, x_j \in C_j} (d(x_i, x_j)) < \min_{x_{k_1} \in C_k} (\min_{x_{k_2} \in C_k} (d(x_{k_1}, x_{k_2}))$ alors les classes i et j ne pourront être séparées que si la classe k est scindée en singletons. Ce cas extrême peut arriver dans des cas de très grandes disparités

6. Perspectives

Du point de vue de la modélisation d'une liaison du type $y = f(x_1, ..., x_n)$ avec f continue, la connexité de $(X_1, ..., X_n)$ est nécessaire, dans le cas contraire des raccordements seront à envisager. De plus la non-connexitéé de l'espace $(Y, X_1, ..., X_n)$ mettra en défaut l'existence de la fonction unique f continue.

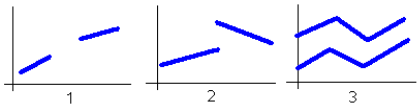

Figure 4. *Exemples pour l'application à la modélisation, dans le premier cas l'espace des X n'est pas connexe dans les autres c'est l'espace (X,Y)*

3. Bibliographie

[BAL 65] BALL G., HALL D., *ISODATA a novel method of a data analysis and pattern classification*, rapport, 1965, Stanford Research Institute.

[SAL 96]., SALEMBIER P., OLIVERAS A., "practical extension of connected operators" *Mathematical Morphology and its application to image and signal processing*, Kluwer, 1998, p 191-206

[WEM 99] WEMMERT C., GANCARSKII J., KORCZAK J ," Un système de raffinement non-supervisé d'un ensemble de hiérarchie de classes ", *Sebag*, 1999, p. 153-160.