

Représentativité et graphe de représentants : une approche inspirée de la théorie du choix social pour la fouille de données relationnelles

Frédéric Blanchard*, Cyril de Runz*
Herman Akdag**, Michel Herbin*

*CRESTIC,
IUT de Reims, Rue des crayères, BP 1025,
51687 REIMS cedex 2
frederic.blanchard@univ-reims.fr,
<http://crestic.univ-reims.fr/>
**LIP6
4 place Jussieu
75252 Paris cedex 05

Résumé. Après avoir défini la *représentativité* dans un ensemble de données relationnelles, nous utilisons cette notion pour construire un *graphe des représentants*. Ce graphe permet de faire émerger des structures arborescentes et des regroupements possibles des données en partitions. Le calcul de la représentativité et la construction du graphe ne requièrent qu'une relation entre les paires d'objets. Notre concept est particulièrement adaptée aux données complexes pour lesquelles on ne dispose pas d'hypothèse a priori. Nous proposons une application directe de notre outil en classification automatique de données complexes.

1 Introduction

Dans la démocratie athénienne, le peuple est réuni sur l'agora et la participation au débat est directe. Le développement des réseaux sociaux et des applications sur internet conduit à réunir sur une agora virtuelle un grand nombre de participants à une activité. Un débat nécessite alors le regroupement à la volée des opinions exprimées, les plus semblables ou similaires sont diffusées par l'intermédiaire d'un représentant. Ces représentants sont fugaces, ils changent et sont désignés à chaque mouvement d'opinion. Ils participent directement au débat mais ne représentent qu'un regroupement d'opinion à un instant donné.

Le travail présenté dans ce papier est une contribution à la détermination de ces représentants dans un flot de données complexes ou d'opinions exprimées. Pour exposer ce travail, nous avons repris le schéma classique de la classification des données.

Nous n'avons pas trouvé, dans la littérature, de définition formelle de la représentativité. Si les statisticiens emploient ce terme lorsqu'ils évoquent la représentativité d'un échantillon, le

sens est toutefois très éloigné de celui que nous évoquions plus haut. En théorie des catégories, la notion de *typicalité* se rapproche de celle que nous proposons. Le degré de *typicalité* (voir Lesot (2006) ainsi que Lesot et al. (2007) et Rifqi (1996)) est défini sur le principe suivant : « un objet est d'autant plus typique qu'il ressemble beaucoup aux membres de sa classe et qu'il est très différent des membres des autres classes ». La différence majeure avec la notion que nous mettons en place est que dans cette approche, le partitionnement en classe doit précéder le calcul du degré de *typicalité*. Le degré de *typicalité* est donc complètement lié à la classification. Dans notre concept, aucune hypothèse n'est faite sur l'ensemble des données, ni sur l'éventuel espace contenant ces données. Le degré de représentativité dont nous proposons la définition n'est calculé qu'à partir des relations entre les données. Elle est complètement indépendante du problème de classification. Elle ne requiert, pour être mise en place, qu'une relation entre les objets pris deux à deux. Notre définition est donc particulièrement adaptée à l'analyse des données complexes, pour lesquelles on ne dispose généralement pas de métrique simple.

Notre définition repose dans son interprétation sur des idées empruntées à la théorie du choix social ce qui permet de la rendre plus « expressive » (nous avons dans un précédent travail (voir Blanchard et al. (2010)), proposé une définition de la représentativité, pour des données numériques, reposant la théorie des ensembles flous, mais dont l'interprétabilité n'était pas aisée). Formellement, elle utilise des outils mathématiques simples et robustes. On peut décomposer son calcul en trois étapes :

1. Expression des préférences individuelles des objets : les objets à étudier sont présentés à travers les relations valuées entre les objets, deux à deux.
2. Transformation des préférences individuelles : les relations sont transformées en rangs, et les rangs en scores de rangs (scores de rangs de Borda, voir Chamberlin et Courant (1983)).
3. Calcul du degré de représentativité par agrégation des préférences individuelles : les scores de rangs obtenus par chaque objet sont agrégés.

La deuxième partie de notre contribution consiste à exploiter ce degré de représentativité pour faire émerger une structuration des objets étudiés. Nous définissons sur ces objets un *graphe des représentants* de la manière suivante :

1. Les degrés de représentativité de tous les objets sont calculés.
2. On détermine les voisinages des objets au sens des k -plus proches voisins (en utilisant la relation pour définir les proximités).
3. On associe à chaque objet, celui, dans son voisinage, dont le degré de représentativité est le plus élevé. Ce deuxième objet est en quelque sorte le « représentant local » du premier.

On obtient ainsi un graphe dont chaque composante connexe est un arbre et possède un unique meilleur représentant (le sommet qui n'a pas de successeur).

Nous proposons enfin, comme conséquence immédiate, une application à la classification automatique de données. En effet, le partitionnement de l'ensemble des objets induit par celui en composantes connexes du graphe des représentants, constitue une classification de l'ensemble initial.

Dans la suite de ce document, nous décrivons toutes les étapes qui permettent de construire le degré de représentativité, puis le graphe des représentants. Nous proposons ensuite une application à la classification automatique. Enfin, nous terminons par les traditionnelles conclusions et perspectives.

2 Degré de représentativité

La première étape de notre méthode consiste à définir la *représentativité* de chaque donnée dans son ensemble initial, et de calculer son *degré de représentativité*. D'un point de vue mathématique, ce calcul repose sur la transformation en rangs des dissimilarités entre les objets (données) considérés, puis sur une agrégation de ces rangs pour produire un indice quantifiant « combien » chaque objet est représentatif de son ensemble.

2.1 Les données relationnelles

Comme nous l'avons déjà expliqué, nous nous intéressons dans ce travail à l'analyse des données relationnelles. On suppose ainsi que l'on dispose d'un ensemble $\mathcal{O} = \{O_1, O_2, \dots, O_n\}$ de n objets à étudier. Contrairement au cas classique où ces objets sont décrits par des vecteurs numériques (caractérisant les « mesures » observées des différentes variables sur ces objets), les données relationnelles sont constituées de valeurs numériques quantifiant la *relation* entre les objets pris deux à deux. On représente ainsi ces données sous la forme d'une matrice $n \times n$:

$$R_{\mathcal{O}} = \begin{pmatrix} R_{1,1} & \cdots & R_{1,n} \\ \vdots & \ddots & \vdots \\ R_{n,1} & \cdots & R_{n,n} \end{pmatrix}$$

où $R_{i,j}$ quantifie la relation entre l'objet O_i et l'objet O_j .

Nous supposons, dans la suite de ce document, que la relation quantifie un éloignement, une différence, entre les objets. Les mesures de *dissimilarité* et de *distance* sont les exemples les plus typiques de ce genre de relation.

2.2 Scores de rangs

La première étape consiste à transformer cette matrice de dissimilarité en matrice de rangs. La transformation en rang est une technique de prétraitement des données qui confère au processus qui l'utilise une robustesse et qui permet d'échapper aux hypothèses sur la distribution des données (voir Friedman (1937)) et de limiter l'impact d'éventuelles données aberrantes. Pratiquement, cette opération consiste à transformer chaque colonne R^i de la matrice $R_{\mathcal{O}}$ en remplaçant chaque valeur de cette colonne par son rang dans l'ensemble trié des valeurs. On obtient ainsi une matrice $n \times n$:

$$RG_{\mathcal{O}} = \begin{pmatrix} RG_{1,1} & \cdots & RG_{1,n} \\ \vdots & \ddots & \vdots \\ RG_{n,1} & \cdots & RG_{n,n} \end{pmatrix}$$

où :

$$RG_{i,j} = 1 + \sum_{k=1}^n \mathbf{1}_{]-\infty; R_{i,j}](R_{k,j})$$

En empruntant le langage de la théorie du choix social, l'opération précédente peut sembler plus intuitive. Considérons un objet quelconque O_j de \mathcal{O} . On peut voir la colonne RG^j comme étant le classement effectué par O_j sur l'ensemble des objets de \mathcal{O} , en fonction de ses préférences, ou de sa ressemblance avec les autres objets. Ainsi $RG_{i,j} = k$ signifie que O_i est le $k^{\text{ième}}$ objet préféré par O_j . Naturellement, O_j est l'objet préféré de O_j .

À l'issue de cette étape, chaque objet classe tous les autres en fonction de ses préférences, et par conséquent, chaque objet se voit classé par les n objets de \mathcal{O} . Ainsi, chaque objet peut être caractérisé par l'ensemble des rangs qu'il obtient dans les n classements. Pour un objet O_i ces rangs sont contenus dans la $i^{\text{ème}}$ ligne de $RG_{\mathcal{O}}$.

L'étape suivante consiste ensuite à agréger les scores correspondant aux rangs obtenus afin d'obtenir un *score global* pour chaque objet, que nous appellerons *degré de représentativité*.

2.3 Degré de représentativité

Pour agréger les « préférences » des objets nous avons choisi d'utiliser la méthode de Borda. Cette méthode consiste à transformer chaque rang en un score puis à les agréger en les sommant. Un objet classé 1^{er} par un autre reçoit n points. Il reçoit $n - 1$ points lorsqu'il est classé 2^{ème}, $n - k + 1$ lorsqu'il est classé $k^{\text{ième}}$, et 1 seul point lorsqu'il est dernier. Chaque objet O_i reçoit ainsi n scores de rangs. En sommant ces scores on obtient un score global qui, divisé par n , définit ce que nous appelons l'indice de représentativité de l'objet O_i dans \mathcal{O} (remarque : diviser par n n'est pas toujours utile ; on préférera, par exemple pour mieux discriminer les objets, conserver la somme sans effectuer cette normalisation).

On a donc :

$$DR(O_i) = \frac{1}{n} \sum_{j=1}^n (n - RG_{i,j} + 1)$$

et donc :

$$DR(O_i) = n - \frac{1}{n} \sum_{j=1}^n \sum_{k=1}^n \mathbf{1}_{]-\infty; R_{i,j}](R_{k,j})$$

Cette notion, comme son nom l'indique, quantifie combien un objet représentatif de son ensemble. En observant la façon dont elle a été construite, on peut facilement vérifier que *plus un objet est « préféré » par les autres, plus son indice de représentativité est élevé (et inversement)*.

2.4 Exemple

Pour illustrer le processus de calcul de l'indice de représentativité, nous allons l'utiliser sur un exemple simple.

On considère les points de \mathbb{R}^2 suivants :

$$\begin{aligned} A &= (1.00, 2.00) \\ B &= (1.00, 3.00) \\ C &= (2.00, 2.00) \\ D &= (3.00, 4.00) \\ E &= (1.00, 0.00) \\ F &= (3.00, 1.00) \end{aligned}$$

dont les relations deux à deux sont induites par la distance euclidienne et représentées dans la matrice de dissimilarité suivante :

R	A	B	C	D	E	F
A	0.00	1.00	1.00	2.83	2.00	2.24
B	1.00	0.00	1.41	2.24	3.00	2.83
C	1.00	1.41	0.00	2.24	2.24	1.41
D	2.83	2.24	2.24	0.00	4.47	3.00
E	2.00	3.00	2.24	4.47	0.00	2.24
F	2.24	2.83	1.41	3.00	2.24	0.00

La transformation des relations en rangs, puis en scores de rangs donnent :

RG	A	B	C	D	E	F	Sco.	A	B	C	D	E	F
A	1	2	2	4	2	3	A	6	5	5	3	5	4
B	2	1	3	2	5	5	B	5	6	4	5	2	2
C	2	3	1	2	3	2	C	5	4	6	5	4	5
D	6	4	5	1	6	6	D	1	3	2	6	1	1
E	4	6	5	6	1	3	E	3	1	2	1	6	4
F	5	5	3	5	3	1	F	2	2	4	2	4	6

Enfin, les scores sont agrégés (en lignes) pour former les degrés de représentativité (sans normalisation) :

	DR
A	28
B	24
C	29
D	14
E	17
F	20

En représentant graphiquement les points dans le plan et en affectant un niveau de gris d'autant plus sombre que le point est représentatif, on obtient la figure 1.

3 Graphe de représentativité

Dans cette partie, nous allons exploiter le degré de représentativité pour construire, sur l'ensemble des objets étudiés, un graphe que nous appelons *graphe des représentants*.

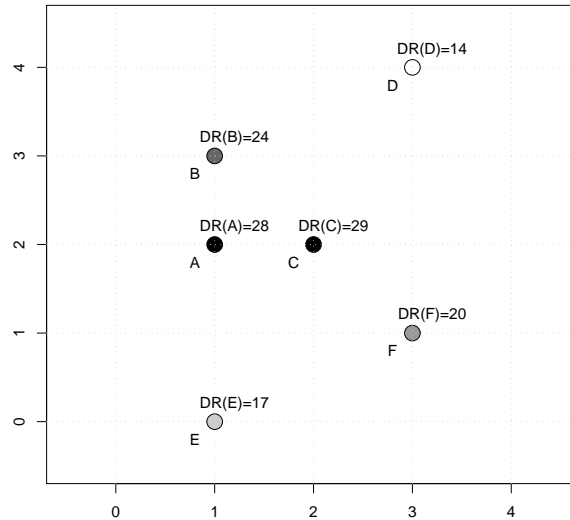


FIG. 1 – Calcul de la représentativité sur un ensemble de six points de \mathbb{R}^2 .

3.1 Construction du graphe

Le principe est d'associer, à chaque objet, son voisin le plus représentatif. Le voisinage d'un objet est déterminé par les k objets qu'il préfère. Autrement dit, le voisinage considéré est un voisinage au sens des k plus proches voisins (k -ppv), dans lequel la notion de proximité est définie grâce à la relation initiale entre les objets.

- On obtient ainsi un graphe orienté $G_k = (X, U)$ (avec $k \in \llbracket 1, n \rrbracket$) dont :
- l'ensemble des sommets X est l'ensemble des objets étudiés (\mathcal{O}) ;
 - l'ensemble des arcs U est défini de la manière suivante :

$$\forall (i, j) \in \llbracket 1, n \rrbracket^2, \quad (O_i, O_j) \in U \Leftrightarrow O_j = \underset{O_p \in V_k(O_i)}{\operatorname{argmax}} (DR(O_p))$$

où $V_k(O_i)$ est l'ensemble des k -ppv de O_i : $V_k(O_i) = \{O_p \in \mathcal{O} / RG_{p,i} \leq k\}$

Cette définition permet à un objet d'être son propre représentant, lorsqu'il est le plus représentatif parmi son propre entourage. D'autre part, le graphe obtenu est une forêt. En effet, il n'est pas nécessairement connexe, mais ne peut contenir ni cycle ni circuit.

La construction de ce graphe entraîne une autre propriété intéressante : chaque composante connexe du graphe contient un et un seul *puits* (un puits est un sommet ne possédant aucun successeur). Ces sommets jouent un rôle particulier dans le graphe. Ils ont en effet la particularité d'être leur propre meilleur représentant et d'être, par transitivité, les meilleurs représentants de

leurs composantes connexes respectives.

3.2 Composantes connexes

Le nombre de composantes connexes est directement lié au paramètre k utilisé pour déterminer les voisinages des objets. Plus k est élevé et plus le nombre de composantes connexes diminue. Lorsque $k = n$ (avec $n = \text{Card}(\mathcal{O})$), le graphe est connexe. Lorsque $k = 1$, chaque objet est son propre meilleur représentant, et le graphe possède n composantes connexes.

Cette relation est illustrée dans l'exemple de la partie suivante (figure 4).

3.3 Exemple

Nous appliquons maintenant notre méthode à un exemple simple pour en illustrer les principaux points.

Considérons un ensemble $\mathcal{O} = \{O_1, O_2, \dots, O_{20}\}$ constitué de 20 points de \mathbb{R}^2 . Cet ensemble est constitué de deux classes $\mathcal{C}_1 = \{O_1, O_2, \dots, O_{10}\}$ et $\mathcal{C}_2 = \{O_{11}, O_{12}, \dots, O_{20}\}$. La figure 2 représente l'ensemble des points dans le plan. Chaque objet est représenté par un disque dont le niveau de gris est d'autant plus foncé que le degré de représentativité est élevé.

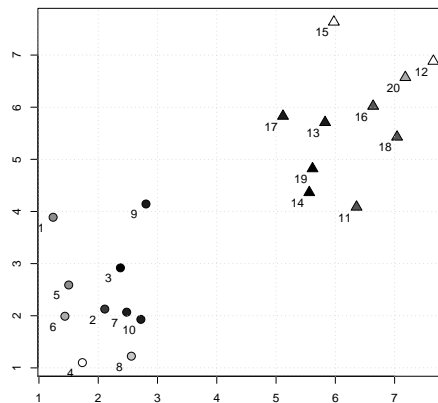


FIG. 2 – Points en 2D : les données et leur degré de représentativité (le niveau de gris est d'autant plus sombre que le degré de représentativité est élevé)

Enfin, la figure 3 illustre le graphe des représentants sur l'ensemble de points.

Représentativité et graphe de représentants d'un ensemble de données relationnelles

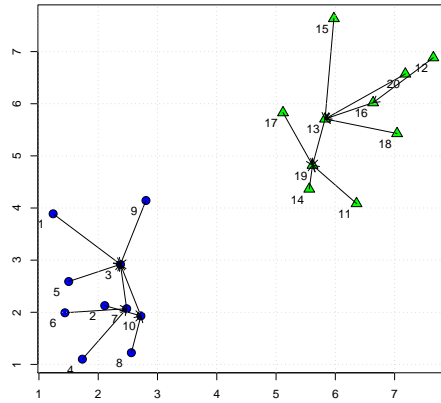


FIG. 3 – Points en 2D : les données et le graphe des représentants

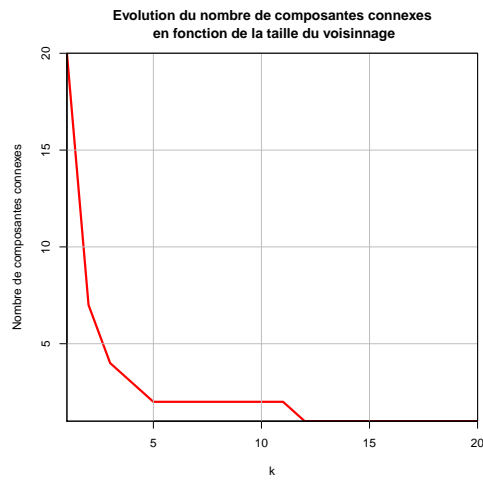


FIG. 4 – Nombre de composantes connexes Vs. k

4 Application à la classification automatique

4.1 Partitionnement

Le calcul de la représentativité et du graphe des représentants associé présente deux intérêts immédiats en classification automatique. Tout d'abord la partition du graphe de représentants permet de façon évidente de partitionner l'ensemble des objets étudiés en classes. En faisant varier le paramètre k (la « taille » du voisinage), on obtient un partitionnement plus ou moins fin. Une conséquence triviale de ce qui a été dit avant est que *plus k est petit et plus le partitionnement est fin (i.e. plus le nombre de classe est élevé)*.

La classification effectuée ainsi possède un certain nombre d'avantages hérités des méthodes impliquées dans le calcul de la représentativité.

Le premier avantage est lié à la nature des données traitées. Les données relationnelles sont un cas plus général que celui où l'on dispose de descriptions vectorielles des objets à étudier, et qui est généralement le pré-requis des algorithmes classiques.

Ensuite, la transformation préalable des dissimilarités en rangs est un outil souvent utilisé par les statisticiens pour obtenir des méthodes non paramétriques. Dans notre contexte, elle permet d'effectuer une classification sans faire d'hypothèse sur la distribution des données. Par ailleurs aucune hypothèse n'est -même implicitement- faite sur la forme des classes. La transformation par rangs offre aussi une robustesse vis-à-vis des valeurs aberrantes.

Enfin, la simplicité des outils théoriques impliqués ainsi que les champs sémantiques des domaines auxquels ils sont empruntés, offrent des possibilités de compréhension, d'interprétation et d'explication que ne permettent pas la plupart des algorithmes.

Remarque : Le choix de la meilleure valeur du paramètre k pour la classification n'est pas aisée. Empiriquement, nous avons déterminé que le choix optimal du paramètre se situait après le « coude », lorsque l'on observe la courbe du nombre de composantes connexes en fonction de k . Cette plage correspond aux valeurs de k pour lesquelles on atteint le premier plateau de stabilisation du nombre de composantes connexes.

4.2 « Meilleurs » représentants

Contrairement aux méthodes de classification comme celles de la famille des k -means, notre méthode de classification ne calcule pas des centres de classes « virtuels » (obtenus par moyenne par exemple), mais extraits, parmi les objets initiaux, des représentants « réels ». Là encore, les calculs sont effectués directement sur les données, à partir de leurs relations deux à deux, et pas dans un hypothétique espace sous-jacent. Cet aspect présente un intérêt non négligeable, notamment dans les situations où l'on souhaite analyser plus précisément les classes à travers leurs représentants.

4.3 Exemple sur des données synthétiques

Nous illustrons maintenant ce processus sur deux jeux de données synthétiques simulées. Les résultats sont illustrés sur la figure 5.

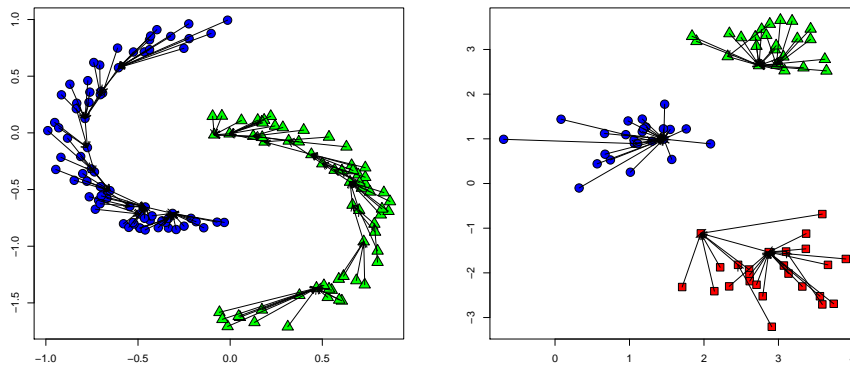


FIG. 5 – Application à la classification automatique

5 Conclusion et perspectives

Nous avons proposé, dans ce papier, une définition de la représentativité sur un ensemble de données relationnelles. Cette notion permet de quantifier « combien » un objet est représentatif de son ensemble. Le calcul du degré de représentativité s'effectue à partir des relations valuées entre les objets, deux à deux. Cet indice est robuste et permet facilement d'être interprété grâce à l'utilisation de la sémantique de la théorie du choix social à laquelle nous empruntons des outils pour effectuer le calcul. Par ailleurs, son calcul ne nécessite aucun a priori sur la distribution des données ou sur l'espace qui les contient.

Nous avons utilisé le degré de représentativité pour construire un graphe des représentants. Ce graphe permet de faire émerger une structuration des objets, en associant à chacun d'entre eux, « celui qui le représente le mieux » parmi « ceux qu'il préfère ».

L'utilisation de ces deux concepts en classification est immédiat et assez intuitif. Nous avons par ailleurs déjà utilisé le concept de représentativité, dans une version préliminaire, pour l'extraction d'éléments représentatifs en archéologie, (de Runz et al., 2008).

L'utilisation de ce travail pour l'analyse des réseaux sociaux nous semble particulièrement opportune et nous envisageons donc d'orienter les applications dans ce domaine. Sur le plan théorique, nous pensons pouvoir exploiter l'aspect hiérarchique du graphe des représentants et l'utiliser pour développer une méthode hybride de classification que nous comparerons aux approches classiques.

Références

- Blanchard, F., P. Vautrot, H. Akdag, et M. Herbin (2010). Data representativeness based on fuzzy set theory. *Journal of Uncertain Systems* 4(3), 216–228.
- Chamberlin, J. R. et P. N. Courant (1983). Representative deliberations and representative decisions : Proportional representation and the borda rule. *The American Political Science Review* 77(3), pp. 718–733.
- de Runz, C., F. Blanchard, E. Desjardin, et M. Herbin (2008). Fouilles archéologiques : à la recherche d'éléments représentatifs. In *Atelier Fouilles de Données Complexes - Conférence Extraction et Gestion des Connaissances - AFDC@EGC*, Sophia Antipolis, France, pp. 95–103.
- Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *J. Am. Statist. Ass.* 32, 675–701.
- Hataway, R. J. et J. C. Bezdek (2003). Visual cluster validity for prototype generator clustering models. *Pattern Recognition Letters* 24, 1563–1569.
- Hataway, R. J., J. C. Bezdek, et J. W. Davenport (1996). On relational data versions of c-means algorithms. *Pattern Recognition Letters* 17, 607–612.
- Lesot, M.-J. (2006). Typicality-based clustering. *Int. Journal of Information Technology and Intelligent Computing* 1(2), 279–292.
- Lesot, M.-J., M. Rifqi, et B. Bouchon-Meunier (2007). *Fuzzy prototypes: From a cognitive view to a machine learning principle*, Chapter Fuzzy Sets and Their Extensions: Representation, Aggregation and Models, pp. 431–452. Springer.
- Rifqi, M. (1996). Constructing prototypes from large databases. In *Proc. IPMU'96*, pp. 301–306.

Summary

As complex objects are often modeled as relational data, we propose a (complex) data mining approach using relations between objects. After defining data representativeness in a relational dataset, we use it to build a graph of representatives (delegates). This graph makes possible the emergence of tree structures and of data groups forming partitions. Finally, our approach is illustrated with a clustering application.