Une mesure de dissimilarité pour des données complexes

Jean Diatta, David Grosser, Henri Ralambondrainy
IREMIA, Université de la Réunion
15, avenue René Cassin – BP 7151
97 715 Saint-Denis Messag. Cedex 9, France

Résumé

Nous proposons une mesure de dissimilarité basée sur la position et le contenu des valeurs descriptives d'objets. Ces objets sont caractérisés par des variables monovaluées ou multivaluées de type quantitatif ou qualitatif. L'intérêt de cette mesure est montré par une étude comparative avec des mesures existantes et par son utilisation dans une méthode de classification. .

1 Introduction

Les mesures de similarités ou de dissimilarités jouent un rôle fondamental dans les méthodes de classification, de reconnaissance de formes, de raisonnement par cas et, plus généralement, en Analyse des Données. Les distances sur des objets caractérisés par des variables numériques sont bien connues et ont été largement étudiées. Cependant, dans les applications réelles, il est courant de mesurer sur un objet des variables de type quantitatif, qualitatif dont les valeurs peuvent être manquantes, inconnues ou multiples (intervalles numériques ou ensembles finis de valeurs). Pour de telles données, les distances usuelles ne peuvent pas s'appliquer. L'approche habituelle pour contourner cette difficulté est de coder de manière *ad hoc* les valeurs pour faire entrer ces valeurs complexes dans un cadre numérique; ce qui conduit à une perte d'information.

Concernant les dissimilarités sur des données comportant des valeurs de type intervalle, différentes propositions ont été faites, notamment par Ichino [3] [4] et Gowda et Diday [2]. Ces derniers proposent une distance qui fait intervenir la "position", "l'étendue" et le "contenu" des descriptions des objets. Les mesures de dissimilarité proposées par ces auteurs sont efficaces dans la plupart des cas, mais ne distinguent pas certaines situations pourtant très différentes [5]. De plus, elles traitent toutes les valeurs descriptives des variables quantitatives comme des intervalles, et ne prennent donc pas en compte le caractère discret des ensembles finis de valeurs.

Nous proposons dans cet article, une mesure de dissimilarité permettant de comparer de manière très fine des objets caractérisés par des variables aussi bien quantitatives que qualitatives. Les valeurs des variables peuvent être des ensembles finis de valeurs discrètes et/ou d'intervalles. Un objet décrit par de telles valeurs sera appelé, dans la suite, "objet complexe".

Si f désigne l'application qui à chaque objet associe sa description, alors la dissimilarité que nous proposons est construite, selon une approche classique, comme l'image réciproque, par f, d'une mesure de dissimilarité préalablement définie sur l'espace de description des objets.

Après avoir introduit la mesure de dissimilarité, nous la comparons, sur des exemples, aux mesures proposées de Hausdorff, de Gowda et Diday [2] et de Ichino [3], puis nous l'utilisons dans une méthode de classification ascendante s'appuyant sur un opérateur d'union cartésienne.

2 Représentation des objets complexes

Soient X un ensemble de n objets et $A = \{a_1, , a_p\}$ un ensemble de variables. On note $\mathbf{dom}(a)$ le domaine (l'ensemble des valeurs possibles) de la variable a.Le domaine d'une variable quantitative est un sous-ensemble borné de $\mathbb{R}$, celui d'une variable qualitative est un ensemble fini, ordonné ou non. Une valeur réelle r peut être considérée comme l'intervalle $[r, r]$. Il est possible de placer dans un cadre unique les variables quelque soit leur type. Pour cela, on considère une variable a comme une application de X dans l'ensemble $\mathcal{V}_a = \mathcal{P}(dom(a))$ des parties de son domaine. Une valeur *manquante* est ainsi codée par l'ensemble vide, la valeur *inconnue* par le domaine tout entier. Une valeur est dite *précise* si c'est un singleton du domaine, (de type) *intervalle* si c'est un intervalle d'un domaine réel. Nous serons emmenés à considérer des ensembles finis de valeurs. Par abus, nour assimilerons ces ensembles à des valeurs et nous dirons, de manière générale, qu'une valeur est (de type) *ensemble* si c'est un ensemble fini de valeurs. L'espace des objets complexes est : $\mathcal{V} := \prod_{i=1}^{i=p} \mathcal{V}_{a_i}$. Un objet x est un p-uple $v = (a_1(x), , a_p(x))$ de $\mathcal{V}$, noté encore $x = V_x^{a_1} \times \cdots \times V_x^{a_p}$ où $V_x^{a_i} = a_i(x)$, $i = 1p$.

3 Mesure de dissimilarité sur des objets complexes

La définition d'une distance sur des variables hétérogènes pose le problème de la normalisation. Comment procéder pour que chaque variable joue un rôle identique dans la mesure de la distance entre deux objets ? La difficulté est d'autant plus grande que nous nous autorisons la présence simultanée de variables monovaluées et multivaluées de différents types. Pour cela, nous avons considéré une valeur V d'une variable a comme une partie de son domaine. Nous définissons une distance normalisée unique sur chaque

variable à partir d'un indice *taille* $s(V)$ adapté à chaque type de variable. Cet indice et d'autres sont introduits ci-dessous.

3.1 Les indices de base

Dans la suite V désignera une valeur d'une variable a. Supposons que a soit une variable quantitative. Si V est un intervalle alors V_{lb} et V_{ub} représentent la borne supérieure et inférieure de V et $\mathbf{b}(V) = \{V_{lb}, V_{ub}\}$ l'ensemble des points extrémités de l'intervalle V. Si V est de type ensemble, alors $\mathbf{b}(V) = \cup_{E \in V} \mathbf{b}(E)$. Nous définissons la borne inférieure V_{lb} et la borne supérieure V_{ub} de V, comme le minimum et le maximum des bornes inférieures et supérieures de tous les éléments de V, respectivement :

$$V_{lb} = \min_{E \in V} E_{lb} \ ; \ V_{ub} = \max_{E \in V} E_{ub}$$

L'enveloppe convexe de V est tout simplement l'intervalle $\mathbf{co}(V) = [V_{lb}; V_{ub}]$. Soit $\mathbf{L}$ la *mesure de Lebesgue* ou en abrégé *L-mesure* sur $\mathbf{dom}(a)$. C'est une fonction réelle positive définie sur $\mathcal{V}_a$ telle que:

$$\mathbf{L}(V) = \begin{cases} V_{ub} - V_{lb} & \text{si } V \text{ est précis ou un intervalle;} \\ \sum_{E \in V} \mathbf{L}(E) & \text{si } V \text{ est de type ensemble} \end{cases}$$

Soient $n_a = \max\{|V_x^a| : x \in X, V_x^a \text{ est de type ensemble}\}$ le *plus grand cardinal* des valeurs de a de type ensemble et $\lambda_a = \min\{\mathbf{L}(V_x^a) : x \in X, V_x^a \text{ est de type intervalle}\}$ *la plus petite L-mesure* des valeurs de a de type intervalle. Le paramètre λ_a est égal à 1 si a ne comporte pas de valeur de type intervalle. Alors on définit la *longueur* de V par

$$\mathbf{l}(V) = \begin{cases} 2\,\mathbf{L}(V) & \text{si } V \text{ est type intervalle;} \\ \mathbf{L}(V) + \frac{\lambda_a |V| \mathbf{L}(\mathbf{co}(V))}{2\, n_a \mathbf{L}(\mathbf{dom}(a))} & \text{si } V \text{ est de type ensemble} \end{cases}$$

Ces indices ont été définis pour mesurer le "contenu" des diverses valeurs. La mesure de Lebesgue aurait été un bon candidat si nous ne traitions que des valeurs de type intervalle. Mais cette mesure n'est pas adaptée pour les ensembles discrets sur lesquels elle est indifféremment nulle. L-mesure $\mathbf{L}(co(V)))$ de son enveloppe convexe qui est normalisé par $\mathbf{L}(dom(a))$ Malgré cette normalisation, nous ne voulons pas qu'une valeur de type ensemble aient une plus grande étendue qu'un intervalle, si une valeur intervalle est présente dans les données. D'où la normalisation par par la L-mesure de la plus petite valeur λ_a pour tenir compte de la présence éventuelle d'une valeur intervalle. de distinguer entre les valeurs $\{b, [c; d], e\}$, $\{b, c, e\}$ et $\{[b; c], d, e\}$. C'est pourquoi, pour une valeur de type ensemble, on ajoute la L-mesure de ses elements, et la longueur d'un intervalle est égale à deux fois sa L-mesure.

Maintenant nous définissons la *taille* d'une valeur V par:

$$\mathbf{s}(V) = \begin{cases} \mathbf{l}(V) & \text{si } a \text{ est quantitative} \\ |V| & \text{sinon} \end{cases}$$

3.2 Définition de la mesure de dissimilarité

Nous définissons la dissimilarité entre deux objets x et y par

$$D(x,y) = \left(\sum_{i=1}^{i=p} \beta_i \, d^r(V_x^{a_i}, V_y^{a_i}) \right)^{\frac{1}{r}},$$

où $r \geq 1$ et où $\beta_i \geq 0$ est le poids de la variable a_i, $i = 1,,p$. Dans la pratique, la valeur du paramètre r sera souvent 1, 2 ou ∞. Pour une variable a, la dissimilarité $d(V_x^a, V_y^a)$ est définie comme la combinaison convexe de deux composants d_P et d_C relatifs à la *position* et *l'étendue*.

$$d(V_x^a, V_y^a) = \eta \, d_P(V_x^a, V_y^a) + \zeta \, d_C(V_x^a, V_y^a)$$

où $\eta, \zeta \geq 0$, $\eta + \zeta = 1$.

3.2.1 Dissimilarité basée sur la position

La position relative de deux valeurs est mesurée à l'aide de la dissimilarité d_P calculée sur les valeurs réelles moyennes:

$$d_P(V_x^a, V_y^a) := \frac{1}{\mathbf{s}(\mathbf{dom}(a))} \left| \frac{1}{|\mathbf{b}(V_x^a)|} \sum_{\alpha \in \mathbf{b}(V_x^a)} \alpha - \frac{1}{|\mathbf{b}(V_y^a)|} \sum_{\alpha \in \mathbf{b}(V_y^a)} \alpha \right|.$$

valeurs telles que $d_P(V_1, V_2) = 0$. Soient μ_1, μ_2 et μ_3 les moyennes de V_1, V_2 et V_3. Alors clairement, $|\mu_1 - \mu_3| = |\mu_2 - \mu_3|$ et $d_P(V_1, V_3) = d_P(V_2, V_3)$. Notons que d_P satisfait l'*inégalité triangulaire*.

3.2.2 Dissimilarité basée sur le contenu

La composante d_C de la dissimilarité basée sur le contenu est définie comme la moyenne géométrique de deux termes d_{SP} et d_{CP}, relatifs aux parties *specifiques* et à la partie *commune* des valeurs, respectivement.

$$d_C(V_x^a, V_y^a) = \sqrt{d_{SP}(V_x^a, V_y^a) * d_{CP}(V_x^a, V_y^a)}$$

La composante relative aux parties spécifiques est:

$$d_{SP}(V_x^a, V_y^a) := \frac{1}{\mathbf{s}(\mathbf{dom}(a))}\, \mathbf{s}(V_x^a \triangle V_y^a).$$

où $V_x^a \triangle V_y^a$ est la différence symétrique entre les valeurs V_x^a et V_y^a. La composante relative à la partie commune est:

$$d_{CP}(V_x^a, V_y^a) := 1 - \frac{1}{\mathbf{s}(\mathbf{dom}(a))}\, \mathbf{s}(V_x^a \cap V_y^a).$$

d_{CP} est en fait une *pseudo-dissimilarité* car $d_{CP}(V, V)$ peut ne pas être égale à 0 pour certaines valeurs de V. Pour une valeur donnée de $d_{SP}(V_1, V_2)$, plus la partie commune est importante entre V_1 et V_2, plus la valeur de d_C est petite. De même, pour une valeur donnée de $d_{CP}(V_1, V_2)$, plus la partie spécifique est grande entre V_1 et V_2 plus la valeur de d_C est importante. triangulaire.

4 Propriétés de la dissimilarité D et comparaison avec d'autres travaux

Dans cette partie, nous nous intéressons aux propriétés de la dissimilarité D pour deux valeurs de r: $r = \infty$ et $r = 1$. Si $r = \infty$, alors, par commodité, on peut considérer les poids β_i comme β_i^r, puisque $\beta_i \geq 0$, pour $i = 1, , p$ et ainsi, pour $x, y \in X$: $D(x, y) = \max_{i \in \{1,,p\}} \beta_i\, d(V_x^{a_i}, V_y^{a_i})$.

Si $r = 1$, alors pour $x, y \in X$: $D(x, y) = \sum_{i=1}^{i=p} \beta_i\, d(V_x^{a_i}, V_y^{a_i})$.

Pour ces deux valeurs de r, on démontre que la dissimilarité D est propre (i.e $D(x, y) = 0$ implique $x = y$) mais qu'elle ne satisfait pas l'inégalité triangulaire. Notons que le paramètre η (resp. ζ) permet de moduler l'importance à accorder à la position (resp. au contenu).

$D(x, y) = \beta_a\, d(V_x^a, V_y^a)$. Sans perte de généralité, on peut omettre $\beta_a > 0$. Il s'ensuit que que pour tout $x, y \in X$, il existe une variable $a \in A$ telle que $D(x, y) = \eta\, d_P(V_x^a, V_y^a) + \zeta\, d_C(V_x^a, V_y^a)$. Alors $D(x, y) = 0$ ssi $\eta\, d_P(V_x^a, V_y^a) = 0$ et $\zeta\, d_C(V_x^a, V_y^a) = 0$.

poids non nul et (2) $\zeta \neq 0$. variable vérifiant les conditions précédentes et que chaque variable a un poids non nul. Alors pour toute variable $e \neq a$, $\eta\, d_P(V_x^e, V_y^e) = 0$ (si la variable e est quantitative) et $\zeta\, d_C(V_x^e, V_y^e) = 0$ puisque

$$\eta\, d_P(V_x^e, V_y^e) + \zeta\, d_C(V_x^e, V_y^e) \leq \eta\, d_P(V_x^a, V_y^a) + \zeta\, d_C(V_x^a, V_y^a).$$

De plus $\eta\, d_P + \zeta\, d_C$ est définie puisque d_C l'est et $\zeta \neq 0$. Donc pour tout $e \in A$, $V_x^e = V_y^e$ et par suite $x = y$.

sans perte de généralité que $\beta_i = 1$, $i = 1, , p$. Alors $x, y \in X$:

$$D(x,y) = \eta \sum_{i=1}^{i=p} d_P(V_x^{a_i}, V_y^{a_i}) + \zeta \sum_{i=1}^{i=p} d_C(V_x^{a_i}, V_y^{a_i}).$$

Egalité qui peut se réécrire plus implement:

$$D(x,y) = \eta\, D_P(x,y) + \zeta\, D_C(x,y).$$

remarquer que D_C est définie. Par contre D ne vérifie pas l'axiome de l'inégalité triangulaire.

l'étendue et à la position, on peut se limiter à une seule variable quantitative. On a alors:

Nous présentons maintenant une comparaison de la dissimilarité que nous proposons avec celles de Gowda et Diday [2], de Hausdorff et de Ichino [3]. Le tableau ci-après est inspiré de [5] où peut être trouvée une étude comparative des trois dernières mesures de dissimilarité. [3] s'est avérée la plus satisfaisante des trois.

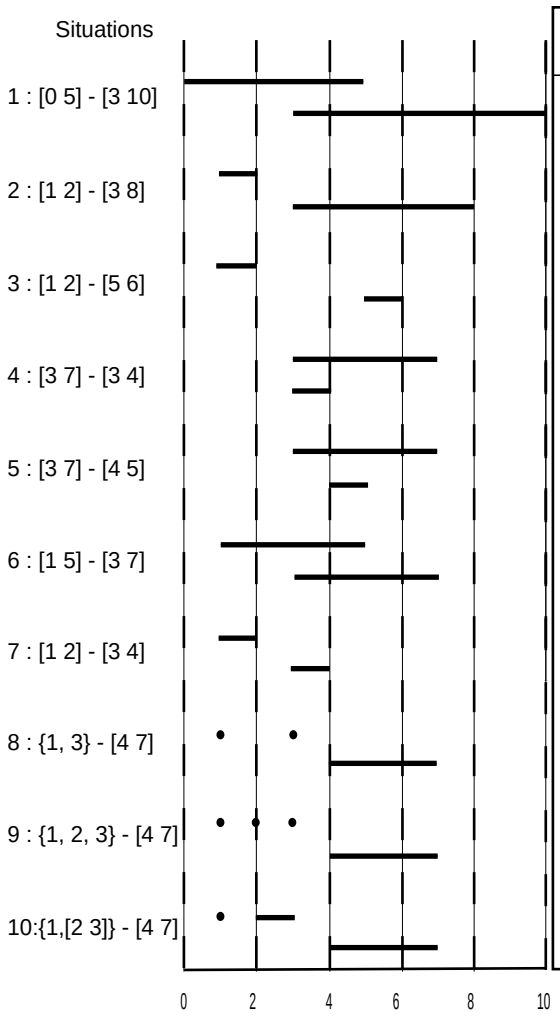
L'expérience porte sur une base de 11 objets décrits par une seule variable quantitative dont les valeurs sont de type intervalle ou ensemble. Il s'agit de voir comment les différentes mesures se comportent en fonction des positions relatives et des étendues respectives des valeurs.

Chaque situation représente les valeurs prises par la variable de description sur deux objets. Les deux valeurs sont séparées par un tiret. En ce qui concerne la mesure que nous proposons, les dissimilarités sont calculées, pour différentes valeurs du paramètre η (trois dernières colonnes), en attribuant un poids unitaire à la variable descriptive.

Dans les situations 4 et 5, les parties commune et spécifiques sont identiques mais les positions relatives des valeurs sont manifestement différentes. Ces deux situations montrent l'insensibilité de la mesure de Ichino par rapport aux positions relatives. Les mesure de Hausdorff, Ichino et Gowda et Diday ne distinguent pas les situations 6 et 7. Ces deux situations sont clairement différentes, étant donné que les deux valeurs de la situation 6 ont une partie commune non vide contrairement à celles de la situation 7.

Dans les situations 8, 9 et 10, le même objet (de description [4; 7]) est respectivement comparé avec trois objets de descriptions différentes. Les mesures de Hausdorff, Gowda et Ichino ne distinguent pas ces trois situations du fait qu'elles traitent les valeurs des variables quantitatives comme leurs enveloppes convexes. Notons que la mesure que nous proposons est sensible à la faible différence entre les situations 8 et 9.

Ces trois dernières situations illustrent bien le fait que la mesure que nous proposons tient compte aussi bien de la cardinalité des valeurs de type ensemble que de l'étendue des éléments de ces valeurs.



	Hausdorff	Gowda	Ichino	DGR (authors) $\eta = 0.3$	$\eta = 0.5$	$\eta = 0.7$
1 : [0 5] - [3 10]	0.5	1.2	0.4	0.519	0.485	0.451
2 : [1 2] - [3 8]	0.6	1.343	0.4	0.506	0.476	0.446
3 : [1 2] - [5 6]	0.4	0.8	0.4	0.345	0.361	0.376
4 : [3 7] - [3 4]	0.3	1.125	0.15	0.304	0.26	0.216
5 : [3 7] - [4 5]	0.2	0.225	0.15	0.274	0.21	0.146
6 : [1 5] - [3 7]	0.2	0.867	0.2	0.343	0.301	0.261
7 : [1 2] - [3 4]	0.2	0.867	0.2	0.283	0.259	0.236
8 : {1, 3} - [4 7]	0.4	1.217	0.35	0.389	0.378	0.366
9 : {1, 2, 3} - [4 7]	0.4	1.217	0.35	0.394	0.381	0.368
10:{1,[2 3]} - [4 7]	0.4	1.217	0.35	0.43	0.407	0.384

FIG. 1: *Comparaison de mesures de dissimilarités*

5 Application

L'algorithme de classification que nous adoptons s'appuie sur l' *opérateur d'union cartésienne*, "$\uplus$", défini, pour tout $x, y \in X$, par $x \uplus y = \left(V_x^{a_1} \cup V_y^{a_1}\right) \times \cdots \times \left(V_x^{a_p} \cup V_y^{a_p}\right)$. Il s'agit d'une méthode ascendante du type de l'algorithme de Ward, où chaque classe devient un nouvel objet, *objet composite*, obtenu par union cartésienne des objets qui la composent. Par analogie avec l'algorithme de Ward, cet objet composite joue le rôle du centre de gravité de la classe, le poids de cette dernière étant le nombre d'objets qui la composent.

Pour illustrer notre approche, nous l'appliquons sur des données relatives à des graisses et des huiles [2], [3]. Sur celles-ci sont mesurées quatre variables quantitatives (*Specific gravity, freezing point, io. value and sa. value*) dont les valeurs sont des intervalles, et une variable qualitative (*m. f. acids*). (Ln), *oleic acid* (O), *palmitic acid* (P) *myristic acid* (M) *searic acid* (S) *arachic acid* (A)*capric acid* (C) and *lauric acid* (Lu).

label	sample name	sp. gravity (g/cm^3)	fr.pt (^oC)	io. value	sa. value	m.f. acids
0	Linseed oil	0.930–0.935	-27 to -8	170–204	118–196	L,Ln,O,P,M
1	Perilla oil	0.930–0.937	-5 to -4	192–208	188–197	L,Ln,O,P,S
2	Cotton-seed	0.916–0.918	-6 to -1	99–113	189–198	L,O,P,M,S
3	Sesame oil	0.920–0.926	-6 to -4	104–116	187–193	L,O,P,S,A
4	Camellia	0.916–0.917	-21 to -15	80–82	189–193	L,O
5	Olive oil	0.914–0.919	0 to 6	79–90	187–196	L,O,P,S
6	Beef-tallow	0.860–0.870	30 to 38	40–48	190–199	O,P,M,S,C
7	Lard	0.858–0.864	22 to 32	53–77	190–202	L,O,P,M,S,Lu

Table 1 : Données sur des graisses et huiles

Le nombre de classes de la partition est déterminé par l'étude de l'histogramme des valeurs de l'indice d'agrégation de la méthode hiérarchique. Un saut significatif est constaté pour un nombre de classes égal à 3.

Remarquons que les objets composites formés à chaque étape de l'algorithme sont bien l'union cartésienne des objets qui les composent, même si les classes (objets composites finaux) sont affichées avec les valeurs de variables quantitatives remplacées par leurs enveloppes convexes. Par exemple, la classe 2 est composée de deux graisses: *beef-tallow* et *Lard*, la valeur d'indice d'iode est comprise entre 40 et 77 (le calcul de la dissimilarité est faite sur la paire d'intervalles $\{[40;48],[53;77]\}$). Pour ces observations, chaque acide de { O,P,M,S,C,L,Lu } est présent dans au moins une de ces graisses .

Cluster	samples	sp. gr.	fr.pt (^oC)	io. value	sa. value	m.f. acids
1	0,1	0.930 – 0.937	-27 to -4	170 – 208	118–197	L,Ln,O,P,M,S
2	6,7	0.858 – 0.870	22 to 38	40 – 77	190–202	L,O,P,M,S,C,Lu
3	2,3,4,5	0.914 – 0.926	-21 to 6	79 – 116	187–198	L,O,P,M,S,A

Table 1 :Description des classes de la partition

6 Conclusion

Dans cet article, nous avons proposé une distance qui présente les avantages suivants:

- traitement dans un cadre unique de variables de différents types,
- possibilité de moduler l'importance de la position et du contenu des valeurs des valeurs descriptives,
- différenciation de situations que les autres approches ne distinguent pas.

Les valeurs inconnues et manquantes sont traitées naturellement du fait que dans notre approche, un attribut est considéré comme étant une application à valeurs dans l'ensemble des parties de son domaine. Cette dissimilarité a été implantée dans le système IKBS [1] pour la classification d'objets complexes.

Références

[1] N. Conruyt, D. Grosser, H. Ralambondrainy, IKBS: An Interative Knowledge Base System for improving description, classification and identification of biological objects, *JICAA97 Ingéniérie des connaissances en Science de la vie: application à la systématique des coraux de Mascareigne* knowledge organisation, *IEEE Trans. Pattern Anal. Mach. Intell.* 7, 592-598, 1985.

[2] K. C. Gowda and E. Diday, Symbolic clustering using a new dissimilarité measure, *Pattern Recognition* 24 : 567-578, 1991.

[3] M. Ichino, General metrics for mixed features : the cartesian space theory for pattern recognition, *IEEE Intl Conf. Syst. Man Cybern.* 14-17, 1986.

[4] M. Ichino, Patern classification based on the cartesian join system : a general tool for feature selection, *IEEE Intl Conf. Syst. Man Cybern.*, 1988.

[5] G. Polaillon, *Organisation et interprétation par les treillis de Galois de données de type multivalué, intervalle ou histogramme*, thèse de doctorat de l'Université Parix IX-Dauphine, 1998. construction of classifications : conceptual clustering versus numerical taxonomy, *IEEE Trans. Pattern Anal. Mach. Intell.* 5 : 396-410, 1983.