Chapitre 3 Visualisation des données complexes

Résumé. Ce chapitre rappelle la problématique de la visualisation dans le cadre de l'ECD et introduit les éléments à prendre en compte lors de la visualisation de données complexes. Dans ce contexte, il nous semble nécessaire d'offrir d'une part une représentation globale définissant des relations particulières entre individus et d'autre part une visualisation locale définissant le contexte d'un individu particulier. Les arbres phylogénétiques et les graphes de voisinages apportent des solutions pour la représentation globale, alors que pour la définition du contexte, nous proposons une carte contextuelle basée sur les voisins d'un individu définie par un graphe de voisinage et permettant une navigation de voisin en voisin. Nous discutons également des problèmes de représentations planes pour des espaces multidimensionnels.

Cependant, pour l'utilisateur, l'important est le contenu des objets complexes. Il doit donc disposer des outils qui lui permettent d'accéder à ces contenus car ce sont ceux-ci qui doivent être traités pour en extraire la quintessence. Ces outils liés à ceux de la représentation des individus définissent notre environnement d'exploration visuelle que nous présentons ici.

Mots clefs : Visualisation, arbres additifs, phylogénie, graphe de voisinage, conservation des proximités, exploration, interaction.

Chapitre 3 – Visualisation de données complexes

3.1 Problématique de la visualisation

L'objectif de la visualisation est de représenter les individus dans l'espace des descripteurs. Dans un contexte d'extraction de connaissances à partir de données, nous y recherchons des structures, caractéristiques, motifs, tendances, anomalies et des relations entre les individus (GRINSTEIN et WARD, 2002). La visualisation apporte une valeur ajoutée souvent due à un cheminement informel et non démontrable (PRYKE, 1996). Cette valeur ajoutée est définie selon FAYYAD *et al.* (2002) comme étant l'augmentation des capacités de perception de l'humain en fournissant une vision perspicace des données. Ainsi, les systèmes basés sur la détection de motifs peuvent être très méthodologiques en ne laissant rien échapper, et effectuer des découvertes statistiquement vérifiables, mais sont pour la plupart dépourvus de la flexibilité humaine.

La taille des graphiques est un point clef de la visualisation et de sa compréhension (HERMAN *et al.*, 2000). En effet, les graphiques très développés posent plusieurs problèmes. Si le nombre d'individus est volumineux, cela peut compromettre les performances en terme de temps de calcul (FEKETE, 2003), de passage à l'échelle des méthodes employées (MENESES et GRINSTEIN, 2002), ou même atteindre les limites de l'affichage (HEALEY, 1996 ; KEIM et KRIEGEL, 1996). En supposant qu'il soit possible d'organiser et d'afficher l'ensemble de ces individus, l'utilisateur pourrait difficilement discerner les individus les uns des autres en raison de leur densité. L'intérêt de la représentation devient alors caduque. En outre, la haute dimensionalité des jeux de données (e.g. on parle couramment de centaines de descripteurs dans l'analyse de documents textuels) ne permet pas leur visualisation directe. Dès lors, il devient impératif d'utiliser des techniques de déformation de l'espace de représentation.

Par ailleurs, dans le cadre de la problématique des données complexes, la notion de voisinage est importante puisqu'elle permet de définir un contexte pour les individus observés. Cette contrainte motive l'utilisation de modèles hiérarchiques ou basés graphe puisqu'ils visent à mettre en apparence les relations directes ou indirectes entre les individus grâce aux arêtes les reliant. De plus, pour une plus grande lisibilité de la représentation des relations de voisinage, il est nécessaire d'effectuer le moins de déformations possibles au sens des positions relatives des individus dans $\mathbb{R}^p$.

En ce sens, nous nous sommes basés sur les modèles d'arbres additifs et de graphes de voisinage conférant des propriétés singulières aux relations entre les individus. Nous avons utilisé une représentation conservant au mieux les proximités des individus pour les modèles arborés, et exploré deux

déformations pour les modèles de graphes de voisinage : la projection de $\mathbb{R}^p$ dans un espace réduit à l'issue d'une analyse factorielle d'une part, et la représentation du voisinage d'un individu d'autre part.

Dans ce chapitre, nous allons ainsi aborder en section 3.2 l'étude de méthodes de représentation. Nous allons décrire dans un premier temps (3.2.1) une méthode issue de la bio-informatique dont ses propriétés singulières mêlées à l'usage d'une méthode de représentation conservant « au mieux » les proximités entre individus vérifient partiellement nos contraintes liées au traitement des données complexes. Puis, dans un second temps (3.2.2), nous reparlerons des graphes de voisinage, mais cette fois-ci dans un contexte de visualisation. Afin de contourner des problèmes de représentation d'un espace de dimension supérieur à 2, nous allons rappeler brièvement le principe de l'ACP pour réduire l'espace d'origine. Nous proposerons également une interface d'exploration locale que nous avons mis en place en utilisant les propriétés des graphes de voisinage.

Enfin, l'ensemble de ces éléments est intégré dans un environnement d'exploration visuelle que nous avons élaboré. Nous présenterons en section 3.3 ses caractéristiques et fonctionnalités en reprenant l'exemple d'un dossier patient.

3.2 Méthodes de représentation

3.2.1 Arbre additif et représentation des proximités

3.2.1.1 Introduction

Les arbres additifs sont des méthodes usuelles des statistiques descriptives multivariées. Ils sont utilisés pour représenter des individus comme feuilles au sein d'un arbre, tels que les proximités deux à deux de ces individus soient représentées « au mieux ». Nous les différencions des arbres de classification hiérarchique (CHANDON et PINSON, 1981) qui ont pour objectif de regrouper en classes homogènes les individus selon un principe d'inclusion et ne peuvent donc pas rendre compte de la proximité deux à deux des individus.

Après un rappel de définitions formalisant les notions que nous allons utiliser, nous allons décrire une famille particulière d'arbres additifs : les arbres phylogénétiques. Nous étudierons une méthode per-

mettant la reconstruction de tels arbres qui est largement utilisée en phylogénie et utilisable pour des individus représentés sous forme vectorielle.

3.2.1.2 Définitions et notations

En continuation de 2.3.2, nous introduisons les concepts que nous allons utiliser par la suite.

Degré. Dans un graphe $G=(\Sigma,A)$, le degré d'un sommet $\alpha\in\Sigma$ est égal au nombre de ses voisins et est noté $d(\alpha)=|\mathcal{V}_\alpha|$.

Feuille, nœud. Tout sommet de degré 1 est appelé une feuille. Les autres sommets sont appelés nœuds.

Arbre étoile, centre. Un arbre étoile est un arbre qui ne possède qu'un seul nœud appelé centre.

Arbre planté, racine. Un arbre planté est un couple (H,r) formé d'un arbre $H=(\Sigma,A)$ et d'un nœud $r\in\Sigma$, appelé racine.

X-arbre, étiquetage, ensemble des étiquettes. Un X-arbre est un couple (H,Ett) formé d'un arbre $H=(\Sigma,A)$ et d'une fonction *Ett* de *X* dans Σ telle que $\forall\alpha\in\Sigma-Ett(X),d(\alpha)\geq 3$. La fonction *Ett* est appelée étiquetage du X-arbre et *X* est l'ensemble des étiquettes et est fini.

Sommet réel, sommet latent. Les sommets dans $Ett(X)$ sont appelés sommets réels et les sommets de $\Sigma-Ett(X)$ sont appelés sommets latents.

X-arbre planté. Un X-arbre planté est un triplet (H,Ett,r) formé d'un X-arbre (H,Ett) et d'un arbre planté (H,r) tel que $\forall\alpha\in\Sigma-Ett(X)-\{r\},d(\alpha)\geq 3$.

X-arbre libre. Un X-arbre (H,Ett) est dit libre lorsque $Ett(X)$ est l'ensemble des feuilles de *H*.

X-arbre additif. Un X-arbre additif (ou valué) est un triplet (H,Ett,L) où (H,Ett) est un X-arbre et (H,L) est un arbre valué.

Distance additive. Soit un arbre valué (H,L) où $H=(\Sigma,A)$. La distance additive entre deux sommets $\alpha,\beta\in\Sigma$, est notée $d_L(\alpha,\beta)$ et est égale à la somme des longueurs des arêtes composant le chemin joignant α à β.

3.2.1.3 Arbre phylogénétique

La phylogénie est une science dont le but est la reconstruction de l'histoire des espèces. Elle suppose une évolution biologique au cours du temps se traduisant par des mutations du code génétique des individus et peut se représenter sous forme arborée. Sa principale contrainte réside au seul accès aux individus observés et non pas aux individus ancestraux. En reprenant le formalisme précédant, les individus observés sont les sommets réels de l'arbre phylogénétique et les individus ancestraux sont les sommets latents. L'arbre phylogénétique est donc un X-arbre où $Ett(X)$ désigne les sommets réels. Nous ferons remarquer qu'un arbre phylogénétique n'est pas forcément un X-arbre libre, i.e. des nœuds peuvent également être des sommets réels si ils sont étiquetés. Du point de vue de l'évolution biologique, ces nœuds étiquetés sont des ancêtres communs observés.

Les arbres phylogénétiques ne sont pas l'apanage des biologistes, citons par exemple son utilisation dans le cadre de l'analyse textuelle (BARTHELEMY et LUONG, 1998). Dans notre travail, nous interprétons les sommets réels comme étant nos individus complexes et le principe de l'évolution comme étant celui du contenu de l'objet complexe et se basant sur les différences de leur représentation vectorielle.

Pour reconstruire de tels arbres à partir des sommets réels, trois familles de méthodes sont principalement utilisées :

1. Les méthodes de parcimonie ;
2. Les méthodes de vraisemblance ;
3. Les méthodes de distances.

Nous n'aborderons pas les deux premières familles puisqu'elles utilisent des données strictement biologiques : les méthodes de parcimonie se basent sur les mutations des gènes et les méthodes de vraisemblance nécessitent la définition d'un modèle d'évolution régissant les mutations de gènes. Les méthodes de distances se basent sur la matrice de dissimilarité $\mathcal{D}$ des individus (les sommets réels) composée des dissimilarités $d(\alpha,\beta) \geq 0, \forall(\alpha,\beta) \in X^2$. Nous utiliserons la notation simplifiée $d_{\alpha\beta}$ lorsqu'il n'y aura pas d'ambiguïté.

Nous pouvons formuler le problème de reconstruction comme étant celui de déterminer la topologie du X-arbre additif $\hat{T} = \left(\hat{H}, Ett, \hat{L}\right)$ avec $\hat{H} = \left(\hat{\Sigma}, \hat{A}\right)$ telle que pour tout couple $(\alpha,\beta) \in X^2$, la distance

additive $d_{\hat{L}}(\alpha,\beta) \geq 0$ soit la plus proche possible, au sens des moindres carrés, de la dissimilarité observée $d(\alpha,\beta)$, ce qui revient à résoudre (3.2.1) :

$$\hat{T}\left(\hat{H}, Ett, \hat{L}\right) = \underset{\hat{H}}{\text{Argmin}}\left(\sum_{(\alpha,\beta)\in X^2} \left[d_{\hat{L}}(\alpha,\beta) - d(\alpha,\beta) \right]^2 \right) \tag{3.2.1}$$

Ce problème est NP-difficile selon (DAY, 1987), cité par (BARTHELEMY et LUONG, 1998), affirmant que la partie difficile consiste à trouver la « bonne topologie » sur laquelle projeter la dissimilarité $\mathcal{D}$. Nous allons ainsi détailler l'heuristique la plus utilisée en reconstruction phylogénétique : la méthode Neighbor-Joining.

3.2.1.4 Méthode de reconstruction Neighbor-Joining

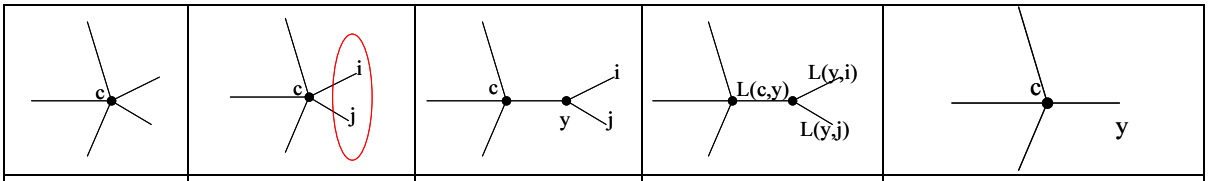
L'heuristique utilisée dans la méthode de Neighbor-Joining (NJ) de SAITOU *et al.* (1987) pour approcher la solution de (3.2.1) est de considérer le X-arbre $\hat{T}$ comme étant libre et de se baser sur le regroupement itératif de la paire de sommets minimisant la longueur totale des arêtes de $\hat{T}$. Ce critère de minimisation est du à l'objectif de respecter le principe d'évolution minimum. De plus, $\hat{T}$ est considéré libre, alors *X* correspond à notre ensemble de *n* individus Ω. Par ailleurs, définir la topologie de $\hat{T}$ nécessite la création d'au plus $(n-2)$ nœuds (cas d'un arbre binaire).

Algorithme 3-1 Neighbor-Joining

```
NJ(entrées: X, 𝒟, n; sorties: Σ, A, L)
Début
      X' ← X; Σ ← X; A ← ø; L ← ø; //Initialisations
      Pour les (n-2) itérations Faire
            a. Créer l'arbre étoile de centre c à partir de X';
            b. Rechercher la meilleure paire de sommets i,j;
            c. Créer le nœud y relié à i, j et c;
            d. MAJ les valuations L(i,y), L(j,y) et L(c,y);
            e. Regrouper (i,j) en y dans 𝒟 selon le lien moyen;
            f. Σ ← Σ+{y}; A ← A+(i,y)+(j,y); X'← X'-{i,j}+{y};
      FinPour
Fin.
```

Dans l'Algorithme 3-1 qui décrit le principe général de cette méthode, la liste des *n* feuilles *X* ainsi que la matrice de dissimilarité *D* sont fournies en paramètres d'entrées et l'algorithme produit le X-

arbre en retournant la liste des sommets Σ, la liste des arêtes A et leur valuation L. Pour une compréhension plus intuitive, le déroulement pas à pas d'une itération est illustré dans la Figure 3-1.



(a) Etoile	(b) Meilleure paire	(c) Ajout de y	(d) Valuations	(e,f) Regroupement et MAJ

Figure 3-1 Déroulement pas à pas d'une itération de l'algorithme NJ

Nous allons maintenant détailler les deux étapes critiques de cet algorithme : la recherche de la meilleure paire de sommets et la mise à jour des valuations (les étapes b. et d. de l'Algorithme 3-1). Nous n'aborderons pas le regroupement puisque le procédé est similaire lors d'une classification hiérarchique, pour plus de détails voir (SAPORTA, 1990, pp. 254-255).

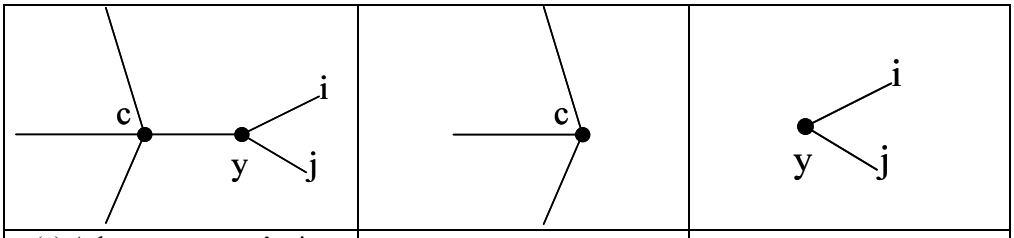
3.2.1.4.1 Recherche de la meilleure paire

La recherche de la meilleure paire (i, j) de sommets se base sur le calcul de la longueur totale des arêtes S_{ij} pour une topologie où i et j sont regroupés en une paire (e.g. Figure 3-1-c) et est calculée à partir de la matrice de dissimilarité $\mathcal{D}$. Cette longueur totale est calculée en adaptant la formule d'un arbre étoile (e.g. Figure 3-1-a) à un arbre avec une arête interne (e.g. Figure 3-1-c). La longueur totale S_0 des arêtes d'un arbre étoile de centre c est alors donnée par l'équation (3.2.2) puisque chaque arête est comptée *(n-1)* fois lorsque toutes les dissimilarités de $\mathcal{D}$ sont ajoutées. En effet, nous rappelons qu'étant dans le cadre de distances additives, alors, la distance $d_{\alpha\beta}$ séparant α et β, deux sommets d'un abre étoile de centre c, est égale à la somme des arêtes formées entre α et β et le centre : $d_{\alpha\beta} = L_{\alpha c} + L_{\beta c}$.

$$S_0 = \sum_{i=1}^{n} L_{ic} = \frac{1}{n-1} \sum_{i<j} d_{ij} \qquad (3.2.2)$$

Par ailleurs, la longueur totale S_{ij} des arêtes pour un arbre comportant une arête interne (c, y) (voir Figure 3-2-a) et un regroupement des feuilles i et j en y, se calcule comme la somme des longueurs S_1

et S_2 des deux arbres étoiles respectivement centrés en *c* et en *y* (Figure 3-2-b et Figure 3-2-c). A la longueur de l'arête interne près, nous obtenons l'équation (3.2.3).



(a) Arbre avec une arête interne *(c,y)* | (b) Arbre étoile centré en *c* | (c) Arbre étoile centré en *y*

Figure 3-2 Décomposition d'un arbre ayant une arête interne

$$S_{ij} = S_1 + S_2 + L_{cy} = \sum_{\substack{k=1 \\ k\neq i, k\neq j}}^{n} L_{kc} + \left(L_{iy} + L_{jy}\right) + L_{cy} \tag{3.2.3}$$

Par ailleurs, la longueur de l'arête interne L_{cy} de la Figure 3-2-a est donnée par l'équation (3.2.4) où le premier terme au sein des crochets correspond à la somme de toutes les dissimilarités incluant (c,y) et les deux autres termes suppriment les longueurs des arêtes dupliquées.

$$L_{cy} = \frac{1}{2(n-2)}\left[\sum_{\substack{k=1 \\ k\neq i;k\neq j}}^{n}\left(d_{ik}+d_{jk}\right) - (n-2)\left(L_{iy}+L_{yj}\right) - 2\sum_{\substack{l=1 \\ l\neq i;l\neq j}}^{n} L_{lc}\right] \tag{3.2.4}$$

Enfin, la somme des longueurs des arêtes (i,y) et (j,y) est égale à la distance arborée $d_L(i,j)$ qui est égale à la dissimilarité d_{ij} :

$$L_{iy} + L_{jy} = d_L(i,j) = d_{ij} \tag{3.2.5}$$

Ainsi, en utilisant les équations (3.2.4) et (3.2.5) dans (3.2.6), nous obtenons la longueur totale S_{ij} en fonction des similarités :

$$S_{ij} = \frac{1}{2(n-2)}\sum_{\substack{k=1 \\ k\neq i;k\neq j}}^{n}\left(d_{ik}+d_{jk}\right) + \frac{1}{2}d_{ij} + \frac{1}{n-2}\sum_{\substack{k<l \\ k,l\neq i;k,l\neq j}}^{n} d_{kl} \tag{3.2.6}$$

3.2.1.4.2 Mise à jour des valuations

A chaque itération, il nous faut mettre à jour les valuations L_{iy}, L_{jy} et L_{cy} correspondant à l'ajout du nœud y au sein de l'arbre (e.g. Figure 3-1-d).

Le calcul de L_{cy} a déjà été défini dans l'équation (3.2.4), en utilisant (3.2.2) et (3.2.5) ; nous pouvons exprimer L_{cy} en fonction des dissimilarités :

$$L_{cy} = \frac{1}{2(n-2)}\left[\sum_{\substack{k=1 \\ k\neq i;k\neq j}}^{n} \left(d_{ik} + d_{jk}\right) - (n-2)d_{ij} - \frac{2}{n-3} \sum_{\substack{k<l \\ k,l\neq i;k,l\neq j}}^{n} d_{kl} \right] \tag{3.2.7}$$

Les valuations L_{iy} et L_{jy} sont calculées à partir de la méthode de FITCH *et al.* (1967) et correspondent aux estimations des moindres carrés de l'arbre de la Figure 3-1-d :

$$\begin{cases} L_{iy} = \left(d_{ij} + d_{iZ} - d_{jZ}\right)/2 \\ L_{jy} = \left(d_{ij} + d_{jZ} - d_{iZ}\right)/2 \end{cases} \tag{3.2.8}$$

$$\text{avec} \begin{cases} d_{iZ} = \dfrac{1}{n-2} \displaystyle\sum_{\substack{k=1 \\ k\neq i;k\neq j}}^{n} d_{ik} \\ d_{jZ} = \dfrac{1}{n-2} \displaystyle\sum_{\substack{k=1 \\ k\neq i;k\neq j}}^{n} d_{jk} \end{cases}$$

Le calcul des valuations étant basé sur la matrice de dissimilarité, la longueur des arêtes n'est pas à interpréter en terme d'évolution rapide ou lente, mais en terme de mutation : plus l'arête est longue, plus les deux sommets sont dissemblants, donc plus deux documents ont un contenu différent.

3.2.1.5 Discussion

La représentation basée sur la construction d'arbre phylogénétique nous semble une méthode pertinente lorsque l'objectif de l'utilisateur nécessite de situer les similarités relatives entre les individus à l'aide d'une représentation plane d'un espace multidimensionnel, par exemple lors de la recherche d'individus aberrants, ou de groupes d'individus similaires.

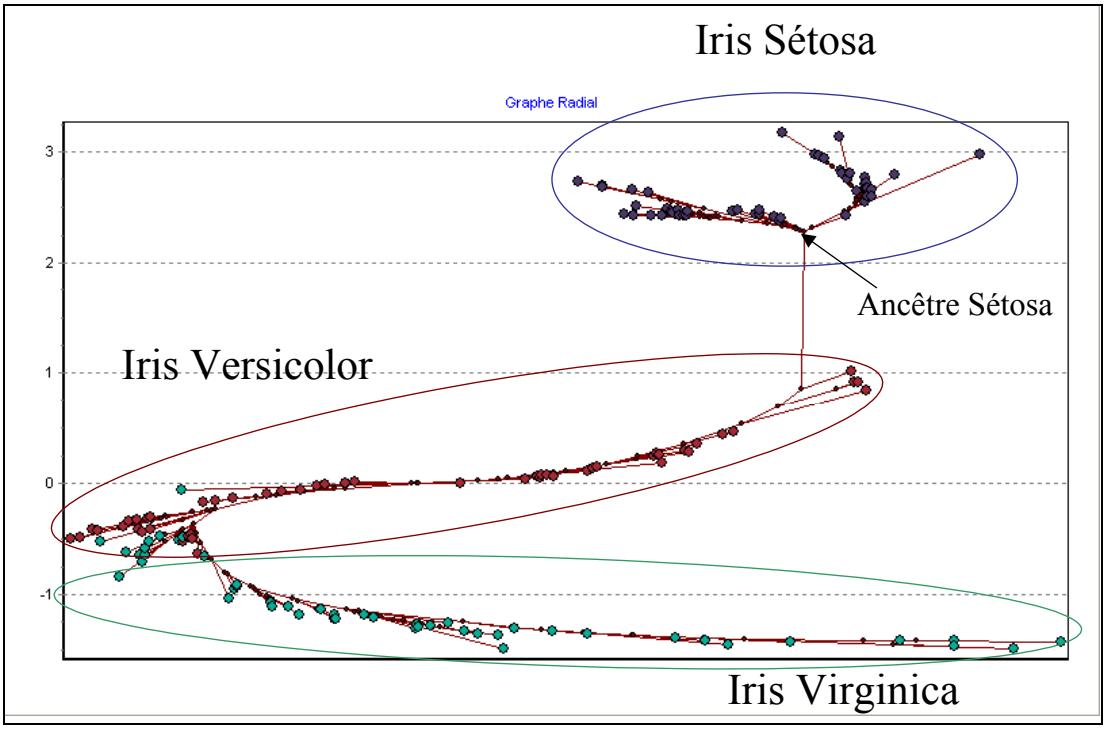


Figure 3-3 Arbre phylogénétique sur le corpus des IRIS

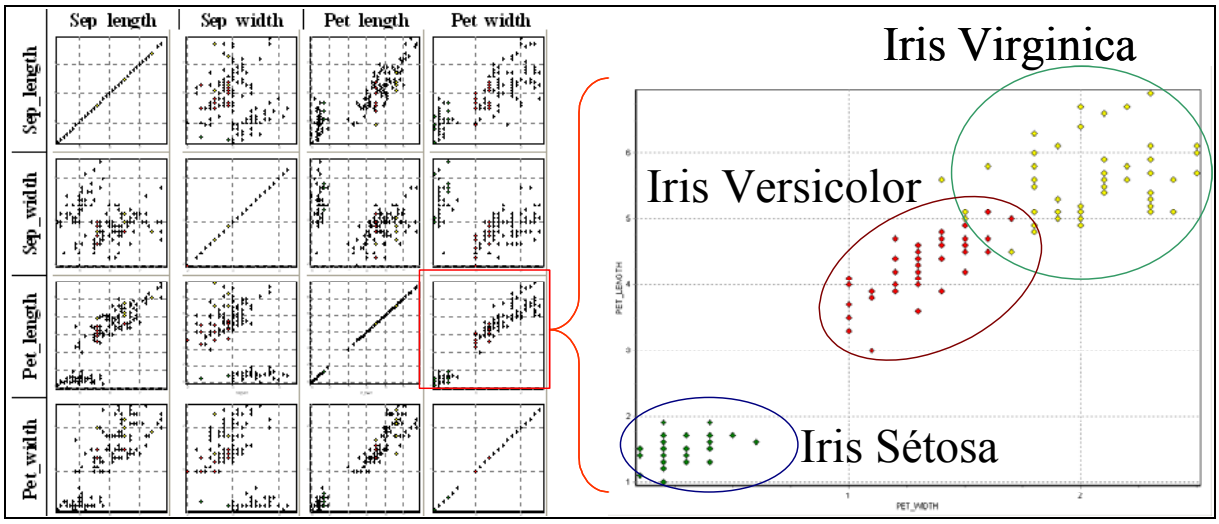


Figure 3-4 Matrice de scatter-plots des IRIS

Pour illustrer ces capacités visuelles, nous avons représenté le jeu de données IRIS (FISHER, 1936) par un arbre phylogénétique (Figure 3-3) et une matrice de scatter-plots (Figure 3-4). Le jeu de données IRIS est composé de 150 individus subdivisés en 3 groupes Iris Sétosa, Virginica et Versicolor. Chacun de ces groupes comprend 50 individus. Enfin, nous connaissons les valeurs de 4 descripteurs indi-

quant la longueur et largeur des sépales et pétales de chacune de ces fleurs. L'objectif de la visualisation sur ce jeu de données peut être de vérifier si cet ensemble de descripteurs permet de différencier les types de fleurs. Nous ferons remarquer que pour faciliter la compréhension cet exemple se limite à l'emploi d'un faible nombre d'individus et de descripteurs.

La lecture de l'arbre phylogénétique est évidente. Nous distinguons clairement les trois familles d'Iris, ainsi qu'un léger empiètement entre les Iris Virginica et Versicolor signifiant une similitude entre ces différents individus. La lecture de la matrice de scatter-plots (Figure 3-4) est plus délicate. Etant composé de 4 descripteurs, le jeu de données est représenté selon toutes les paires de descripteurs possibles. Il nous semble difficile pour l'utilisateur de synthétiser l'ensemble de ces représentations. En étudiant un scatter-plot particulier (selon les variables longueur et largeur des pétales sur la Figure), nous voyons également les 3 familles d'Iris ainsi qu'un empiètement des Virginica et Versicolors. Cependant, l'utilisateur peut se demander si les autres descripteurs vont pouvoir contribuer à différencier ces quelques Iris Virginica des Versicolors. L'exploration de la matrice pour répondre à ces questions est peu aisée puisqu'il faut identifier un individu particulier et le reporter sur les différents scatter-plots, alors que le graphe phylogénétique y répond sans ambiguïtés.

La représentation phylogénétique offre en outre la définition de relations entre des individus et des sommets latents. Nous interprétons ces relations comme des relations d'héritage des sommets latents vers les individus. L'héritage est constitué d'un ensemble de caractéristiques d'un sommet latent plus ou moins partagées par les individus dépendamment de la longueur du plus court chemin les séparant du sommet latent. En reprenant l'exemple des IRIS, nous remarquons sur la Figure 3-3 un ancêtre commun aux Sétosa. Ce dernier peut être considéré comme un individu prototype puisque c'est un individu synthétique et dont les valeurs des caractéristiques sont représentatives des Iris de type Sétosa. Par contre, il est malaisé de définir un ancêtre commun pour les autres types d'Iris. Nous pensons qu'il en existe plusieurs rendant ainsi la lecture plus confuse sur ce point.

La lecture du graphe phylogénétique a pour inconvénient le principe de lecture du graphique. En effet, la dissimilarité entre deux sommets est égale à la longueur du plus court chemin les séparant dans l'arbre et non pas la longueur du segment les reliant directement.

Du point de vue méthodologique, notre choix de méthode de reconstruction s'est porté sur la méthode NJ. Cette méthode a pour avantage d'être en complexité $O(n^3)$ comparée aux autres méthodes telles ADDTREE de SATTAH *et al.* (1977), premier algorithme de reconstruction, en $O(n^4)$, ou encore la

méthode des Groupements de BARTHELEMY *et al.* (1986), présentée dans (BARTHELEMY et GUENOCHE, 1988), ayant une complexité supérieure à $O(n^4)$!

L'inconvénient de NJ et de ADDTREE est de ne produire que des X-arbres libres et binaires, ce qui est évidemment restrictif. La méthode des Groupements supprime ces deux inconvénients, mais hélas au tribut d'un coût prohibitif.

3.2.2 Représentation de graphes de voisinage

3.2.2.1 Introduction

Au Chapitre 2, nous avons présenté des méthodes de construction de graphes de voisinage. Ces derniers ont été utilisés alors dans un cadre de prédiction d'un individu $\alpha \in \Omega$ à partir de ses voisins notés $\mathcal{V}_\alpha$. Ici, nous les utilisons dans un cadre descriptif. L'intérêt de représenter de tels graphes réside dans la mise en avant des relations de proximité et de distance entre individus. Ces relations sont établies à partir de l'ensemble de l'espace de représentation $\mathbb{R}^p$. Cependant, lorsque *p>3*, le problème du choix de l'espace de représentation se pose. En effet, d'une part il est souvent délicat de choisir deux variables particulières parmi les *p* pour définir un plan sur lequel les individus et leurs relations seront projetés. D'autre part, visualiser dans un plan des relations définies dans un espace multidimensionnel implique leur enchevêtrement, rendant très difficile la lecture du graphe.

Pour répondre à ce problème, nous avons utilisé deux stratégies différentes. La première consiste à déterminer un espace de représentation ayant les propriétés suivantes :

- La dimension de cet espace doit fortement être inférieure à celle de l'espace initial afin que la représentation des relations soit plus claire ;
- La position relative de deux individus de ce nouvel espace doit conserver « au mieux » la notion de proximité de l'espace initial.

Par construction, l'Analyse en Composantes Principales (ACP) (PEARSON, 1901) répond à ces critères. Etant largement utilisée, nous rappellerons seulement les principaux points de cette méthode, décrite dans de multiples ouvrages comme (LEBART *et al.*, 2000, pp. 32-66), et les différentes utilisations que nous en avons faites en section 3.2.2.2.

La seconde stratégie que nous avons développé consiste en une représentation locale du voisinage d'un individu. Son principe et sa mise en œuvre seront expliqués en section 3.2.2.3.

3.2.2.2 ACP pour la réduction d'espace

Notre ensemble de données *D* est plongé dans l'espace de représentation $\mathbb{R}^p$. Moyennant des précautions pour rendre homogènes les données, comme par exemple l'opération de centrage-réduction, la matrice *D* définit *n* points dans $\mathbb{R}^p$.

Dans le cas où *D* est une table de contingence, c'est-à-dire où les observations sont des fréquences, alors la distance utilisée pour calculer les distances entre individus est généralement la distance du χ^2. Cette situation apparaît typiquement dans le cas des données textuelles. Dans ce cadre, l'individu *i* de *D* devient $x_i' = \left(x_i^1, \cdots, x_i^p\right) = \left(f_i^1, \cdots, f_i^p\right)$. Pour nous ramener dans le cas général de l'ajustement de $\mathbb{R}^p$ alors nous transformons les *p* coordonnées des individus suivant le changement de variable suivant :

$$\tilde{x}_i' = \left(\frac{f_i^1}{f_i^\bullet \sqrt{f_\bullet^1}}, \cdots, \frac{f_i^p}{f_i^\bullet \sqrt{f_\bullet^p}} \right) \tag{3.2.9}$$

où $f_i^\bullet = \sum_{j \in [1,p]} f_i^j$ sont les effectifs marginaux en lignes et $f_\bullet^j = \sum_{i \in [1,n]} f_i^j$ sont les effectifs marginaux en colonne.

L'objectif de la réduction est de rechercher un sous espace de dimension q où $q \ll p$, tel que $d_q\left(x_i, x_j\right) \approx d_p\left(x_i, x_j\right), \forall i, j \in \Omega$. Les vecteurs propres $U_1, \cdots, U_q$ de $D'D$ seront les vecteurs sur lesquels les individus sont projetés, où D' est la transposée de *D*. Nous faisons remarquer que U_k, $k \in [1,q]$ est le k^{ème} vecteur propre associé à λ_k la k^{ème} plus grande valeur propre de $D'D$, et nous notons $U = \left(U_1, \cdots, U_q\right)$ la matrice composée de ces *q* vecteurs propres.

Dès lors, nous obtenons facilement les nouvelles coordonnées de l'individu *i* de *D* dans $\mathbb{R}^q$ par l'équation (3.2.10) :

$$\widehat{x_i} = x_i' U \tag{3.2.10}$$

Comme le précisent nombres d'auteurs tels (SAPORTA, 1990, p. 178), la réduction de dimension n'est possible que si il existe une certaine redondance entre les p variables. Dans le cas de l'indépendance de ces dernières, l'ACP sera inefficace à réduire la dimension.

L'ACP est présentée ici dans le cadre de la réduction de l'espace afin de pouvoir visualiser notre graphe de voisinage dans les différents plans factoriels. En outre, nous souhaitons l'utiliser pour construire un graphe de voisinage sans croisement d'arêtes. L'utilisation du premier plan factoriel nous apparaît alors comme une approche intéressante pour les raisons déjà évoquées (e.g. conservation « au mieux » des distances).

Néanmoins, il ne faut pas perdre de vue que par construction, l'ACP déforme l'espace d'origine, et qu'à fortiori fixer q à 2 peut modifier fortement la réalité des relations calculées par le graphe de voisinage ; c'est dans ce sens que l'utilisation de la carte contextuelle révélant ces « vraies » relations prend toute son importance. Pour permettre à l'utilisateur de prendre conscience de cette dérive, nous lui proposons l'indicateur usuel du pourcentage de l'inertie expliquée, noté Q, afin de lui indiquer la qualité de la représentation utilisée (variant entre 0 et 100%). Dans le cas de la représentation dans le premier plan factoriel, cet indicateur Q_{1-2} est donné par :

$$Q_{1-2} = \frac{\lambda_1 + \lambda_2}{\sum_{k \in [1,p]} \lambda_k} \tag{3.2.11}$$

3.2.2.3 Carte contextuelle

3.2.2.3.1 Objectif

Cette méthode que nous avons élaborée permet une visualisation simple d'un individu et de ses voisins définis à l'aide d'un graphe de voisinage. L'objectif de la méthode est de permettre de se concentrer sur un individu particulier mais également sur son voisinage directe, souscrivant ainsi à la contextualisation de l'individu et autorisant alors une meilleure interprétation en tenant compte des invariants entre lui et ses voisins.

3.2.2.3.2 Méthode

Cette méthode utilise les résultats issus de l'élaboration du graphe de voisinage construit dans l'espace $\mathbb{R}^p$. Par ailleurs, nous avons choisi de représenter au sein de cette carte contextuelle un maximum de

huit voisins. Ce choix peut être considéré comme arbitraire mais permet d'éviter de surcharger la carte reposant sur une interface simple (cf. 3.2.2.3.3).

L'Algorithme 3-2 produit la liste triée des voisins pour l'élaboration de la carte contextuelle de l'individu *centre* passé en paramètre. La matrice des arêtes *A* est passée en paramètre d'entrée. Elle est de dimension *n*. Une valeur nulle entre deux individus indique qu'il n'existe pas de relation de voisinage entre-eux ; une valeur non nulle indique la dissimilarité séparant les deux voisins. Par convention, nous avons posé $A[i,i]=0,\forall i$.

Algorithme 3-2 Carte contextuelle

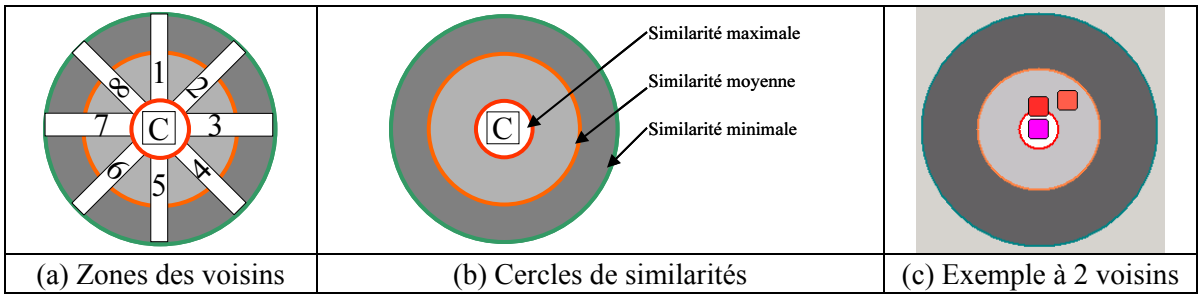
```
CarteContextuelle(entrées: centre, n, A; sortie: V)
Début
    V' ← ø; //Initialisation
    Pour i ← 1 à n Faire  //Ajout trié des voisins de centre
        Si (A[centre,i] ≠ 0) alors
           j ← 1; Fin ← Faux;
           Tant que NON(Fin) Faire
               Si (A[centre,i] < A[centre,V'[j]])
               Alors
                  Inserer(V',i,j); //Insère i à la jème position du tableau V
                  Fin ← Vrai;
               Sinon
                  j ← j+1;
                  Fin ← j>n;
               FinSi
           FinTantQue
        FinSi
    FinPour
    Pour i ← 1 à Min(8,|V|) Faire
        V[i] ← V'[i];
    FinPour
Fin.
```

3.2.2.3.3 Représentation

Par construction, tous les individus de *V* retournés par l'Algorithme 3-2 sont des voisins de l'individu *centre*. La visualisation d'arêtes représentant ces relations est donc inutile dans le cadre de cette carte. Les relations de voisinage entre les individus de *V* bien que potentiellement intéressantes risqueraient de surcharger la lecture du graphique, allant ainsi à l'encontre de notre objectif de simplicité. Nous n'allons donc pas les représenter ; dès lors les positions relatives des voisins les uns des autres n'a pas de signification particulière. L'information qu'il nous reste à mettre en avant est les dissimilarités en-

tre l'individu *centre* et chacun des voisins. Pour ce faire, nous avons placé l'individu *centre* au milieu de la carte, et la position des voisins, en forme de spirale, est assignée en fonction de deux paramètres :

- Le premier paramètre est la position du voisin dans V qui indique la zone où il sera affiché sur la carte (cf. Figure 3-5-a). Le premier voisin sera affiché dans la zone 1 et les suivants dans les zones successives selon la rotation horaire ;
- Le second est la similarité entre l'individu *centre* et un voisin qui permet d'indiquer la distance entre l'individu *centre* et le voisin. Pour une lecture plus aisée, nous avons ajouté sur la carte trois cercles indiquant les similarités singulières (cf. Figure 3-5-b) : maximale (cercle rouge le plus proche du centre), moyenne (cercle intermédiaire orange) et minimale (cercle périphérique vert).



(a) Zones des voisins | (b) Cercles de similarités | (c) Exemple à 2 voisins

Figure 3-5 Carte contextuelle

Remarque : si les similarités entre les individus de Ω ne varient pas dans l'intervalle [0, 1], nous nous y ramenons en prenant pour bornes les similarités extrêmes.

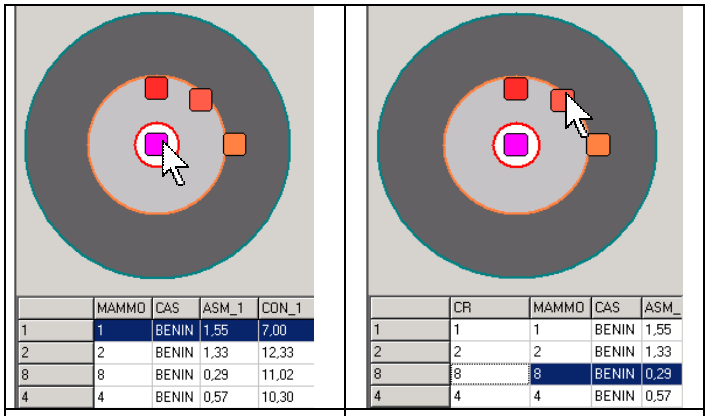
3.2.2.3.4 Interactions

Comme nous l'avons déjà exprimé, cette carte contextuelle vise à permettre l'interprétation d'un individu en fonction de ses voisins. Cela implique la visualisation de ces derniers à partir de cette carte. Travaillant avec des données complexes, il nous faut pouvoir visualiser à la fois des données simples (e.g. numériques, symboliques) et des documents textuels, images vidéos, … Nous avons ainsi intégré à notre carte de contextualisation les deux mécanismes suivants :

- Le premier consiste à surligner les données simples d'un individu de la carte contextuelle rangées sous forme tabulaire sous la carte (cf. Figure 3-6). La sélection de l'individu à mettre en relief se fait lors du passage de la souris sur l'un d'entre eux ;
- Le second consiste à permettre l'édition de documents dans une fenêtre spécifique, s'activant par menu contextuel pour chaque individu de la carte (cf. Figure 3-7). Pour faciliter leur visualisation,

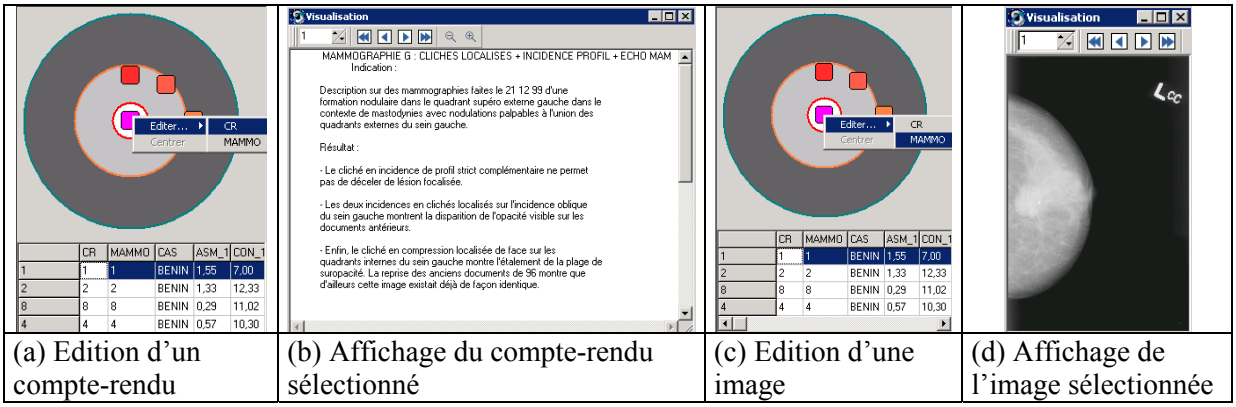
un menu de navigation permet d'afficher le contenu du voisin précédent et du suivant. Cette fenêtre s'adapte bien entendu à la nature du document.

En reprenant l'exemple d'un dossier patient, la carte contextuelle de la Figure 3-6 permet de visualiser rapidement l'ensemble des voisins d'un dossier patient. Le déplacement de la souris au-dessus des divers voisins met en surbrillance les données numériques du dossier voisin concerné, permettant ainsi à l'utilisateur de mieux interpréter les relations de voisinage. Au sein de notre environnement, nous avons privilégié l'accès à la lecture des documents composant l'objet complexe. Comme l'illustre la Figure 3-7, un menu contextuel permet d'éditer le compte-rendu d'un patient ou encore une mammographie.



(a) Individu n°1 (b) Individu n°8

Figure 3-6 Mise en relief de l'individu courant

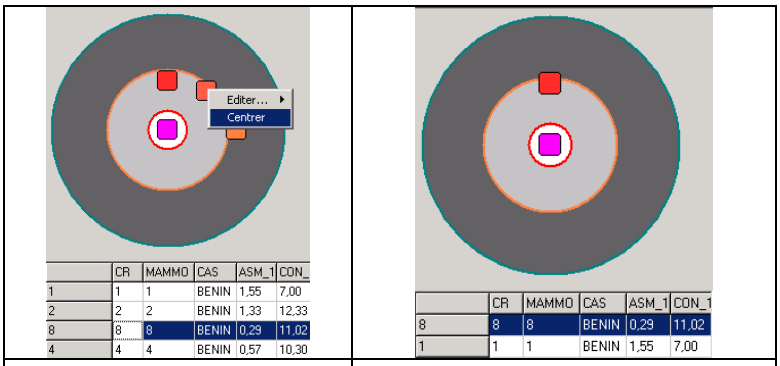


(a) Edition d'un compte-rendu (b) Affichage du compte-rendu sélectionné (c) Edition d'une image (d) Affichage de l'image sélectionnée

Figure 3-7 Edition de documents complexes

3.2.2.3.5 Navigation

Le rôle de la navigation est de permettre de contextualiser n'importe quel individu du voisinage directe de l'individu *centre*. Lors des étapes interactives, il se peut que l'utilisateur s'intéresse plus particulièrement à un voisin plutôt qu'à l'individu *centre*. Dès lors, il doit lui être possible de contextualiser ce voisin particulier. Pour ce faire, nous « centrons » ledit voisin et réappliquons l'Algorithme 3-2. Comme le montre la Figure 3-8-a, cette action est accessible à l'aide d'un menu contextuel.



(a) Choix de l'individu 8 (b) Carte centrée sur l'individu 8

Figure 3-8 Centrage de la carte contextuelle

3.2.2.4 Discussion

Les choix effectués pour cette représentation en forme de spirale peuvent naturellement être discutés et certainement améliorés. Ici, notre objectif était de mettre en place un outil simple et autonome. Notre critère de simplicité doit permettre une interprétation aisée, sans artefact. Le critère d'autonomie vise à autoriser l'accès à toutes les données observées au moyen de cet outil.

Cependant, il nous semble malgré tout intéressant d'évaluer l'apport de l'intégration d'assistants visant à aider à la contextualisation de l'individu *centre*. En prenant l'exemple de documents textuels, un assistant de contextualisation serait de mettre en relief au sein du texte de l'individu *centre* les éléments qu'il aurait en commun avec ses voisins.

3.3 Réalisation d'un environnement d'exploration visuelle

3.3.1 Introduction

Notre objectif, rappelons-le, est la visualisation de données complexes. Dès lors, ce prototype doit permettre au minimum une visualisation aisée de toutes les données, y compris des documents complexes comme des images, des textes, des vidéos. Cependant, dans le cadre de données complexes, il nous semble nécessaire de compléter la visualisation en abordant ce problème sous l'angle des relations de proximité entre individus. Nous avons présenté un panel d'outils permettant de répondre partiellement à ce problème :

- Les arbres additifs permettent une visualisation des proximités des individus deux à deux, mais n'offrent pas une lecture évidente des relations de voisinage ;
- Les graphes de voisinage permettent la définition de ces relations de voisinage dans $\mathbb{R}^p$ mais leur visualisation dans un plan factoriel peut s'avérer délicate en raison de multiples croisements d'arêtes ;
- Les graphes de voisinage construits dans un plan factoriel et projetés dans ce dernier ne contiennent pas de croisements, mais toutes les relations de voisinage ne sont pas vraies dans l'espace d'origine $\mathbb{R}^p$;
- La carte contextuelle permet une représentation des relations de voisinage sans risque de croisements, mais ne s'applique qu'à un seul individu.

Ces outils sont complémentaires dans le sens où chacun d'entre eux apporte une solution différente et que les inconvénients qu'un outil peut apporter sont corrigés par un autre outil. L'outil que nous proposons va donc s'appuyer sur l'intégration et la collaboration de ces outils.

3.3.2 Description de notre proposition

Nous proposons un outil s'articulant sur trois éléments graphiques (voir par exemple Figure 3-10 et Figure 3-11). Le premier élément est une représentation globale du corpus de données complexes. Le choix de la représentation (arbre phylogénétique ou graphe de voisinage) est défini par l'utilisateur. De même, lors du choix d'une représentation par graphe de voisinage (cf. Figure 3-9), l'utilisateur détermine son espace de projection (e.g. premier plan factoriel). Le deuxième élément, situé sous le pre-

mier, se consacre à la visualisation élémentaire des données. Elle est composée d'un tableau affichant les données simples de la totalité du corpus, et de zones d'affichage de données complexes (dans nos exemples, il y a une zone textuelle et une zone image). Le dernier élément, situé à droite des deux premiers, est constitué de la carte contextuelle calculée dans l'espace d'origine et non dans l'espace de projection.

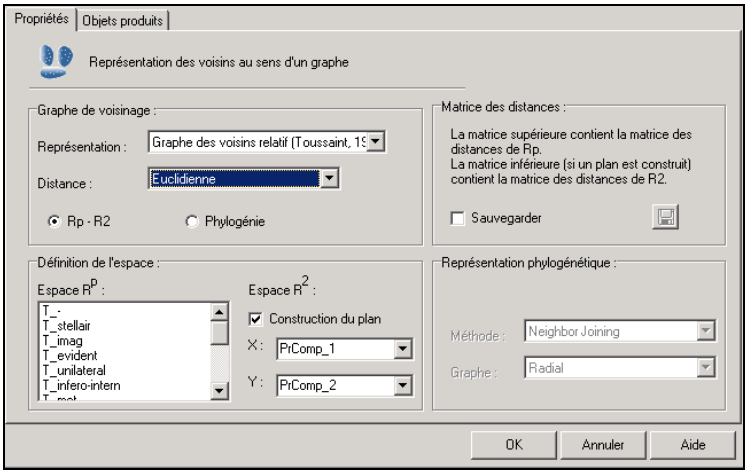


Figure 3-9 Paramètres de la visualisation de données complexes

La cohérence globale de cette interface se fait par la notion d'*individu courant*. L'individu courant est un individu du corpus qui a été sélectionné par l'utilisateur. Cet individu particulier est alors affiché ou mis en évidence dans tous les zones de l'outil :

- Pour la représentation globale, une légende est affichée sous l'individu courant (« BENIN Individu n°5 » sur Figure 3-10) ;
- Pour la visualisation des données, la ligne de cet individu est mise en relief dans le tableau, et les documents complexes de cet individu sont affichés dans les zones prévues à cet effet ;
- La carte contextuelle est centrée sur l'individu courant.

De même que l'individu courant est affiché selon les modalités de chacun des éléments visuels de cet outil, il peut être sélectionné à partir de ces mêmes zones :

- A partir de la représentation globale, un clic sur les individus de la carte le met automatiquement en tant qu'individu courant ;

- A partir de la visualisation tabulaire des données, un double-clic sur la ligne d'un individu le sélectionnera ;
- A partir de la carte contextuelle, c'est l'opération de centrage qui y répond.

Enfin, pour conserver une cohérence de représentation, les voisins de l'individu courant, construits dans l'espace d'origine, sont systématiquement mis en relief dans la représentation globale : l'utilisateur peut visualiser les voisins au sein d'un arbre phylogénétique ou discerner les « vrais » voisins lors d'une construction d'un graphe de voisinage à partir d'un plan factoriel.

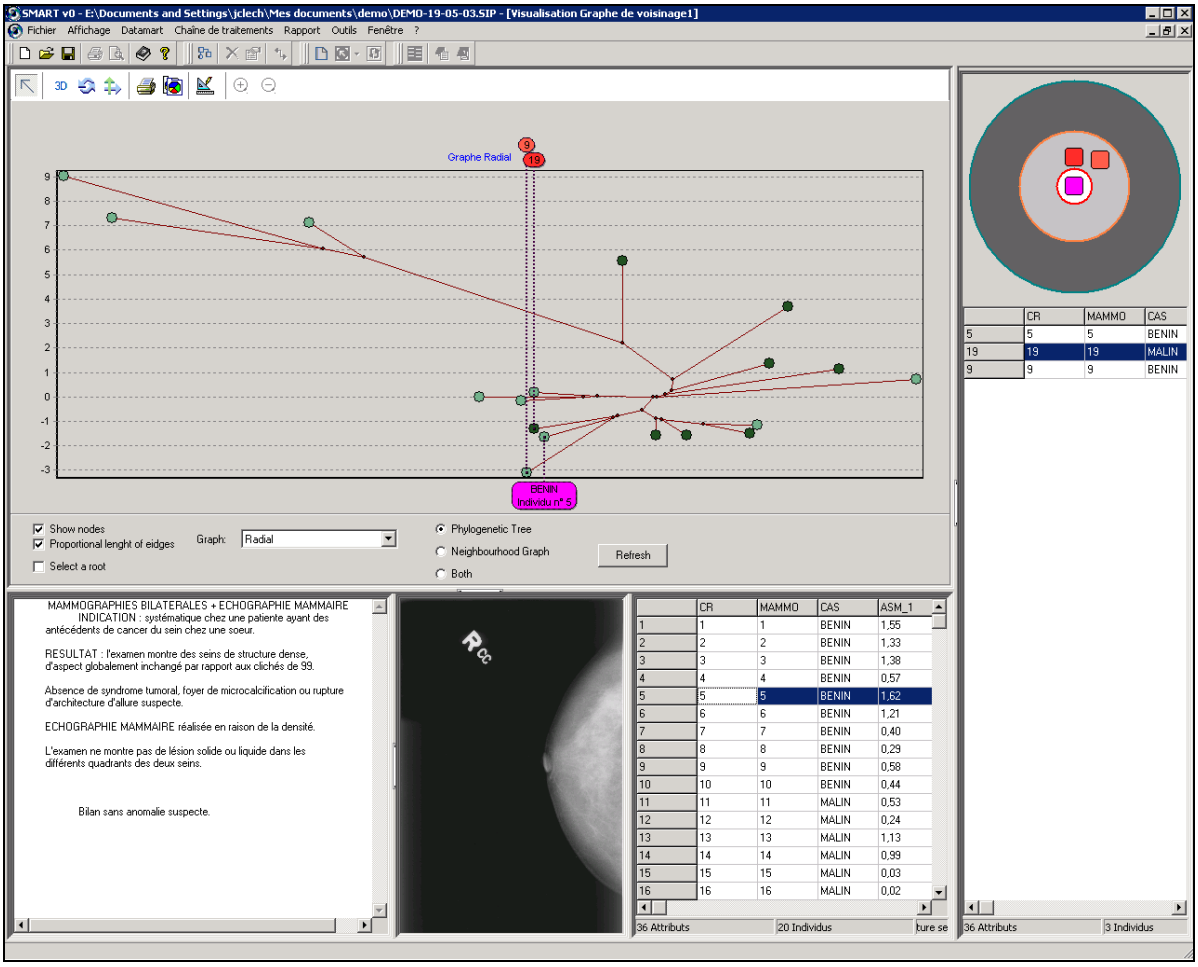


Figure 3-10 Vue arbre phylogénétique de l'outil de visualisation de données complexes

La Figure 3-10 et la Figure 3-11 sont des copies d'écran issues de notre environnement d'exploration visuelle. Le corpus représenté est celui des dossiers patients. L'utilisateur a activé la visualisation globale de l'objet complexe n°5. Le compte-rendu médical et la mammographie faisant parti de ce dossier médical sont alors affichés. Les données numériques de ce dossier sont mis en surbrillance. Nous remarquons qu'il s'agit ici d'un cas bénin. Ce dossier possède 2 voisins, les dossiers 9 et 19. Ils sont affichés au sein de la carte contextuelle où leurs données numériques sont également visibles. Les deux figures diffèrent de part leur représentation globale du corpus : la première représente les relations issues de l'arbre phylogénétique, alors que la seconde représente celles calculées par le graphe de voisins relatifs.

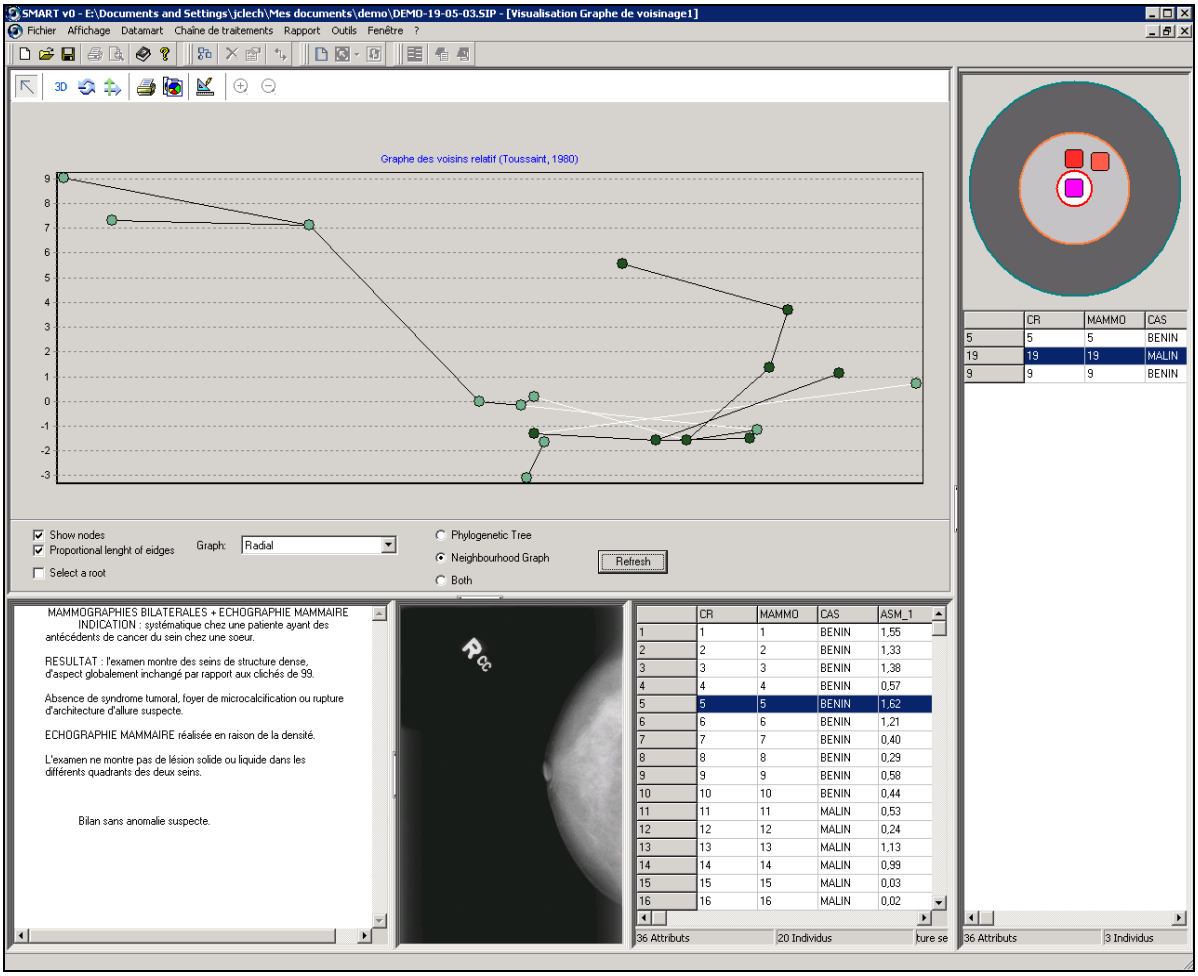


Figure 3-11 Vue graphe de voisinage de l'outil de visualisation de données complexes

3.4 Conclusion

Dans ce chapitre, nous avons décrit diverses approches permettant la représentation de données soit en ayant la propriété de conserver « au mieux » les distances entre individus, soit en permettant la visualisation des relations de voisinage. A partir d'eux, nous avons proposé un outil de visualisation d'objets complexes, dont l'une des spécificités est selon nous la possibilité de visualiser le voisinage d'un individu particulier.

Notre outil ne se résume pas à une simple adjacence des méthodes, mais plutôt à la mise à disposition d'outils globaux et locaux totalement intégrés les uns aux autres notamment de part les interactions implicites entre ces divers éléments.

Cette proposition est un prototype nécessitant nombre d'évolutions, principalement sur le plan de l'interaction homme-machine. Mais outre son existence, il a pour intérêt de nous permettre de valider les spécifications (et d'en définir de nouvelles) pour le développement de méthodes de visualisation des données complexes.