L'analyse factorielle pour la construction de cubes de données complexes*

Riadh Ben Messaoud[♭], Sabine Rabaséda[‡] et Omar Boussaid[‡]
Laboratoire ERIC, Université Lumière Lyon 2
5, avenue Pierre Mendès–France,
69676, Bron Cedex, France
[♭]rbenmessaoud@eric.univ-lyon2.fr
[‡]{sabine.rabaseda | boussaid}@univ-lyon2.fr

Résumé

Les données complexes présentent des difficultés pour leur structuration, leur stockage et leur analyse. En particulier, l'analyse en ligne (OLAP) s'avère inadaptée pour ces données. L'association de l'OLAP et de la fouille de données est une solution possible pour l'analyse des données complexes. La grande dimensionnalité est un des aspects de la complexité des données. Les cubes de données complexes construits avec un très grand nombre de dimensions sont difficiles à analyser et à explorer. Nous proposons une nouvelle approche basée sur les méthodes factorielles pour pallier ce problème. Notre approche permet de créer un nouvel espace de représentation pour un cube de données complexes. Elle consiste à générer à partir d'un cube initial, un nouveau cube avec moins de dimensions. Ceci permet une meilleure visualisation et exploration des données complexes.

1 Introduction

Les données complexes présentent de nos jours un enjeu important pour leur structuration, leur stockage et leur analyse. Cepedant, il n'existe pas de définition unique pour les données complexes. Elle dépend directement du contexte des domaines d'applications (domaine médical, domaine de l'emploi, marketing, etc.) Par exemple, dans celui de l'emploi, le profil d'un candidat peut être considéré soit comme un simple individu statistique décrit par des variables qualitatives et/ou quantitatives ; soit comme un objet complexe décrit en langage naturel dans un document texte ou dans une page HTML. Cet objet complexe véhicule une information sur le profil constitué de plusieurs éléments sous forme de descripteurs. Plus généralement, nous considérons que les données sont complexes lorsqu'elles vérifient au moins une des conditions suivantes[1] :

– elles sont multiformats : l'information est portée par des données de types différents (*données numériques, symboliques, textes, images, sons, vidéos* ...)
– elles sont multistructures : structurées, semi-structurées ou non structurées (*bases de données relationnelles, collection de documents XML, documents textes* ...)
– elles sont multisources : provenant de plusieurs sources (*bases de données, web* ...)
– elles sont multimodales : un même phénomène est décrit par plusieurs canaux ou points de vue (*échelles ou langues différentes* ...)

*Ce travail est financé par la Fondation VideorBis pour la Recherche et l'Emploi sous l'égide de la Fondation de France.

[1]http ://bdd.univ-lyon2.fr/ ?page=donnees_complexes

– elles sont multiversions : les données sont évolutives en termes de définition ou de valeur (*bases de données temporelles, recensements périodiques* ...)

L'entreposage des données complexes est une solution nécessaire pour leur préparation à l'analyse. Dans le contexte des entrepôts, une donnée complexe peut être modélisée en un élément d'observation adapté à l'analyse. On parle d'un "*fait*" OLAP (On-line Analytical Processing). Typiquement, un fait est exprimé par une ou plusieurs mesures et peut être décrit par un ensemble de descripteurs. La prise en compte des données complexes dans le contexte de l'analyse en ligne nous amène à repenser les techniques classiques d'entreposage [2]. Du côté de l'analyse, les outils OLAP actuels s'avèrent inadaptés aux données complexes. Dans [4], nous avons proposé un nouvel opérateur d'agrégation en ligne des "*faits complexes*" selon une approche utilisant une technique de fouille de données. Nous pensons que, dans les années à venir, la combinaison des techniques de la fouille de données et de l'OLAP donnera lieu à une nouvelle génération d'opérateurs d'analyse en ligne. De nombreux travaux ont déjà montré l'intérêt de cette combinaison. Nous renvoyons le lecteur à [4], où nous présentons un survol des différentes approches proposées dans ce domaine.

Nos travaux s'insèrent dans cette démarche générale qui associe la fouille de données à l'analyse en ligne. Nous présentons dans cet article une nouvelle approche pour traiter le problème de réduction de la dimensionnalité des cubes dans le cas des données complexes. Notre objectif consiste à créer un nouvel espace de représentation multidimensionnel des données formé d'un nombre réduit de dimensions. Cet espace offre une meilleure représentation des données et une meilleure adéquation à la visualisation et à la navigation OLAP. Cette approche est basée sur l'utilisation des méthodes de l'analyse factorielle.

Cet article est organisé en trois sections. Dans la première, nous présentons la problématique et les motivations de ce travail. Nous détaillons dans la section suivante les différentes étapes de la démarche que nous proposons. Enfin, nous exposons dans la dernière section quelques éléments de discussion liés à cette nouvelle approche.

2 Motivations

Dans les systèmes décisionnels, les données sont représentées selon un modèle en étoile. Autour d'une table de faits centrale contenant une ou plusieurs mesures à observer, il existe plusieurs tables de dimensions comprenant des descripteurs. Ces tables sont alors représentés dans une structure multidimensionnelle adaptée à l'analyse qu'on appelle les *cubes* de données [1]. Un fait est donc représenté par un ensemble de modalités provenant des dimensions d'un cube et observé par une ou plusieurs mesures ayant des propriétés d'additivité plus ou moins fortes. La vocation de l'OLAP est de fournir à l'utilisateur des opérateurs pour résumer et naviguer dans les données afin d'y découvrir des informations pertinentes. Cependant, la navigation dans les cubes de données ne permet pas souvent de trouver des résultats intéressants. La difficulté de la navigation dans les données s'accroit avec l'augmentation de la dimensionnalité du cube. Les moyens actuels de l'analyse en ligne ne permettent pas d'évaluer les niveaux de pretinence des axes d'analyse dans un cube. Dans le cas des données complexes cette difficulté s'amplifie. A titre d'exemple, pour un cube dont les faits représentent des images, les axes d'analyse possibles s'avèrent très nombreux. En effet, une image peut être décrite par une large gamme de variables tels que les descripteurs sémantiques, visuels ou les métadonnées.

Devant le problème de la dimensionnalité des cubes de données complexes, l'analyse en ligne doit disposer d'outils d'assistance afin de mieux guider l'utilisateur vers les axes répondant au mieux à ses objectifs d'analyse. Ces outils doivent aussi créer de nouveaux espaces de représentation. Ces espaces permettent de produire une meilleure visualisation ajustant au mieux le nuage des faits et mettant en avant des points de vue intéressants pour l'analyse. Pour cela nous envisageons faire recours aux méthodes factorielles afin de parvenir à réduire la dimensionnalité d'un cube. Partant d'un cube initial à d dimensions et à N faits (d-*cube*), nous pouvons générer un nouveau cube à r dimensions (r-*cube*, $r < d$). Les r nouvelles "*dimensions factorielles*" correspondent aux axes fournis par une méthode factorielle. Elles traduisent de fortes relations avec les données qui peuvent être évaluées par les inerties liées aux facteurs (les valeurs propres). En plus, l'analyse factorielle nous permet de donner du sens à chacune de ces nouvelles dimensions en fonction des modalités provenant des dimensions de départ.

3 Démarche

Prenons l'exemple du d-$cube$ d'images de la figure 1(a). Ce cube est constitué de trois dimensions ($d = 3$) et d'une seule mesure ($m = 1$). Les dimensions décrivent la texture des niveaux de gris des images (l'homogénéité ASM, l'entropie ENT et le contraste CON). La mesure correspond à la caractéristique moyenne de couleur $L1Norm$. La figure 1 résume, en quatre étapes, notre démarche pour passer du d-$cube$ au r-$cube$.

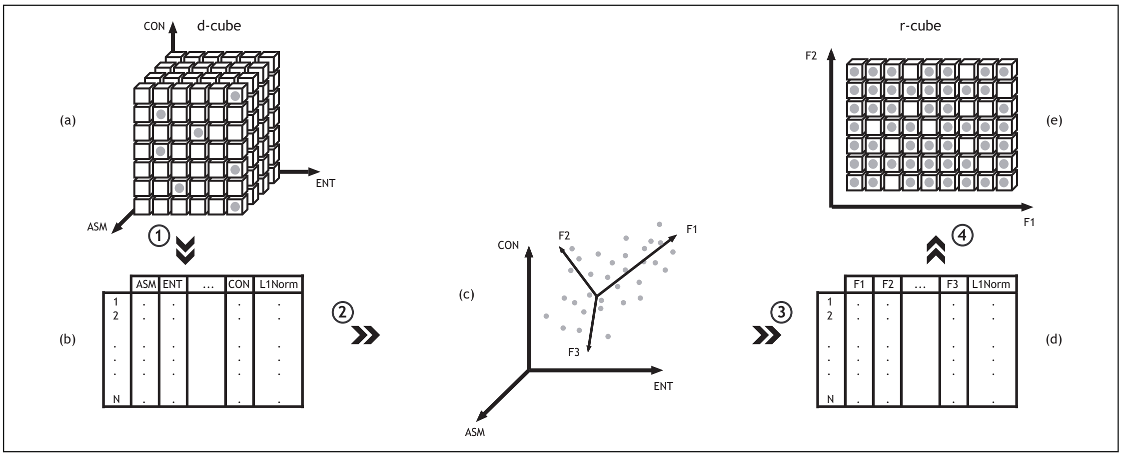

FIG. 1 – Les étapes de construction de la représentation factorielle d'un cube de données

La **première étape** consiste à aplatir le d-$cube$ et à générer un tableau "Individus-Variables" où les faits sont transformés en individus et les dimensions en variables (Figure 1(b)). Les m mesures ne seront pas considérées comme variables mais plutôt comme des poids que nous pouvons affecter aux individus. Classiquement, les dimensions sont formées de modalités symboliques. Par exemple, une dimension géographique contient des noms de villes. Dans ce cas une telle dimension se transforme en une variable qualitative. Dans certains cas, nous pouvons aussi trouver des dimensions ayant des modalités numériques qui se transforment en variables quantitatives.

Dans la **deuxième étape**, selon la nature des variables obtenues, on applique une méthode factorielle à la population des faits. Si les d variables sont qualitatives, on construit le tableau disjonctif complet et on déroule l'analyse des correspondances multiples (ACM). Sinon, on centre et on réduit les variables quantitatives et on déroule l'analyse en composantes principales (ACP). Les deux méthodes permettent la création d'axes factoriels qui représentent au mieux les faits et qui traduisent des relations avec les modalités des dimensions de départ. Chaque axe factoriel (ou facteur) est caractérisé par une valeur propre indiquant la dispersion des faits dans sa direction [3].

Dans la **troisième étape**, nous récupérons les coordonnées des individus fournies par la méthode factorielle. Nous obtenons donc un nouvel espace de représentation pour les faits provenant du d-$cube$ initial. L'application d'une méthode factorielle sur ces faits ne modifie en rien leur propriétés intrinsèques. En effet, un fait garde sa mesure de départ tant que sa position absolue dans l'espace ne change pas. Seule la position relative par rapport aux axes factoriels change. Les m mesures sont donc reliées aux faits dans le nouvel espace (figure 1(d)).

Enfin, la **quatrième étape** consiste à sélectionner les r premiers facteurs qui contribuent le plus à l'inertie totale du nuage des faits. Nous retiendrons par exemple les r premiers axes dont le pourcentage cumulé des inerties est supérieur à 80%. Dans l'exemple de la figure 1(e), nous retenons les deux premiers axes ($r = 2$). Les axes retenus correspondent aux nouvelles *dimensions factorielles* du r-$cube$. Dans le nouveau cube, un fait est caractérisé par les coordonnées de ses projections sur les nouvelles dimensions et par sa mesure initiale. Ainsi, le cube obtenu a un nombre réduit de dimensions ($r < d$) et permet une meilleure représentation des faits.

4 Discussion

La présente approche est une solution possible au problème de la grande dimensionnalité des cubes de données complexes. Ce problème est largement abordé par les méthodes de sélection de variables dans l'apprentissage supervisé. Cependant, dans le contexte d'une analyse exploratoire OLAP, chaque dimension d'un cube peut se révéler importante selon les besoins d'analyse. Nous ne cherchons pas à sélectionner les meilleures dimensions, mais à les résumer et à mieux représenter les faits. Cependant, quelques points restent à résoudre dans notre approche :

- au niveau du choix des r meilleurs facteurs, il est vrai que les pourcentages d'inertie des axes factoriels représentent la part de l'inertie totale prise par chaque facteur. Mais leur appréciation doit aussi tenir compte du nombre d'individus et de variables de départ ;
- la pondération des faits dans la méthode factorielle va affecter l'ordre des distances qui les séparent dans le cube initial. Ceci aura-t-il une incidence sur la représentation de ces faits dans le nouveau cube et par conséquent sur la qualité de l'analyse qu'on peut mener sur ce dernier ?
- une méthode factorielle ne peut traiter que des tableaux homogènes. Toutes les variables sont quantitatives pour l'ACP ou qualitatives pour l'ACM. Ceci présente un problème dans notre approche sachant que dans le cas des données complexes, les dimensions peuvent avoir des modalités numériques et/ou symboliques ;
- enfin, il est nécessaire de construire un indicateur pour apprécier la qualité du cube de données en terme de distribution et de volumétrie. Cet indicateur nous permettra de déduire le gain réalisé avec notre approche en passant du d-$cube$ au r-$cube$.

Références

[1] W. H. INMON, *Building the Data Warehouse*, John Wiley & Sons, (1996).

[2] J. DARMONT AND O. BOUSSAID, *Managing and Processing Complex Data for Decision Support*, Idea Group Publishing, (2005).

[3] L. LEBART AND A. MORINEAU AND M. PIRON, *Statistique Exploratoire Multidimensionnelle*, Dunod, (1995).

[4] R. BEN MESSAOUD AND O. BOUSSAID AND S. RABASEDA, November 2004, *A New OLAP Aggregation Based on the AHC Technique*, in Seventh ACM International Workshop on Data Warehousing and OLAP (DOLAP 2004), pp. 65–72.