Analyse exploratoire par k-Coclustering avec Khiops CoViz

Bruno Guerraz*, Marc Boullé*, Dominique Gay*,
Vincent Lemaire*, Fabrice Clérot*

*Orange Labs
2, avenue Pierre Marzin, F-22307 Lannion Cedex, France
firstname.name@orange.com

Résumé. En analyse exploratoire, l'identification et la visualisation des interactions entre variables dans les grandes bases de données est un défi (Dhillon et al., 2003; Kolda et Sun, 2008). Nous présentons Khiops CoViz, un outil qui permet d'explorer par visualisation les relations importantes entre deux (ou plusieurs) variables, qu'elles soient catégorielles et/ou numériques. La visualisation d'un résultat de coclustering de variables prend la forme d'une grille (ou matrice) dont les dimensions sont partitionnées: les variables catégorielles sont partitionnées en clusters et les variables numériques en intervalles. L'outil permet plusieurs variantes de visualisations à différentes échelles de la grille au moyen de plusieurs critères d'intérêt révélant diverses facettes des relations entre les variables.

1 Khiops CoViz : Visualisation des modèles en grille

Khiops CoViz, développée en Flex, est la brique logicielle de visualisation de Khiops Coclustering (KHC) [1]. Étant données, deux (ou plus) variables catégorielles ou numériques, KHC réalise un partitionnement simultané des variables : les valeurs de variables catégorielles sont groupés en clusters et les variables numériques sont partitionnées en intervalles – ce qui revient à un problème de coclustering. Le produit des partitions uni-variées forme une partition multivariée de l'espace de représentation, i.e., une grille ou matrice de cellules et il représente aussi un estimateur de densité jointe des variables. Afin de choisir la "meilleure" grille M^* (connaissant les données) de l'espace de modèles $\mathcal{M}$, nous exploitons une approche Bayésienne dite Maximum A Posteriori (MAP). KHC explore l'espace de modèles en minimisant un critère Bayésien, appelé $cost$, qui réalise un compromis entre la précision et la robustesse du modèle :

$$cost(M) = -\log(\underbrace{p(M \mid D)}_{\text{posterior}}) = -\log(\underbrace{p(M)}_{\text{prior}} \times \underbrace{p(D \mid M)}_{\text{vraisemblance}}) \quad (1)$$

KHC construit aussi une hiérarchie des parties de chaque dimensions (i.e., clusters ou intervalles adjacents) en utilisant un stratégie agglomérative ascendante, en partant de M^*, la grille optimale résultant de la procédure d'optimisation, jusqu'à $M_\emptyset$, le modèle nul, i.e., la grille (unicellulaire) où aucune dimension n'est partitionnée. Les hiérarchies sont construites en fusionnant les parties qui minimisent l'indice de dissimilarité $\Delta(c_1, c_2) = cost(M_{c_1 \cup c_2}) - cost(M)$,

1. http://www.khiops.com – Pour plus de détails sur l'implémentation de KHC, voir Boullé

où c_1, c_2 sont deux parties d'une partition d'une dimension de la grille M et $M_{c_1 \cup c_2}$ la grille après fusion de c_1 et c_2. De cette manière, la fusion de parties minimise la dégradation du critère $cost$, donc minimise la perte d'information par rapport à la grille M avant fusion. L'utilisateur peut ainsi choisir la granularité de la grille nécessaire à son analyse tout en contrôlant soit le nombre de parties soit le taux d'information (i.e., le pourcentage d'information gardé dans le modèle : $IR(M') = (cost(M') - cost(\mathcal{M}_\emptyset))/(cost(M^*) - cost(M_\emptyset))$. La grille optimale M^* et les hiérarchies correspondantes constituent les principales structures de notre outil de visualisation.

2 Interface utilisateur : Exploration & Interactivité

La figure 1 présente l'interface utilisateur de l'outil pour un sous-ensemble de données de la base DBLP. Nous considérons des données à trois dimensions ($Author \times Year \times Event$) pour 16 000 auteurs ayant publié au moins une fois dans une des principales conférences de Bases de Données ou de Fouille de Données – $Year$ est l'année de la publication. Le panneau principal de visualisation de l'outil est composé des hiérarchies de deux dimensions sélectionnées, des parties terminales des hiérarchies et de la composition d'une partie sélectionnée. La visualisation de la grille correspondante est aussi disponible via le panneau principal (voir la figure 2 à gauche). L'outil permet de naviguer parmi les parties d'une dimension tandis que les deux autres dimensions sont fixées et dédiées à la visualisation (pour le cas à plus de deux variables).

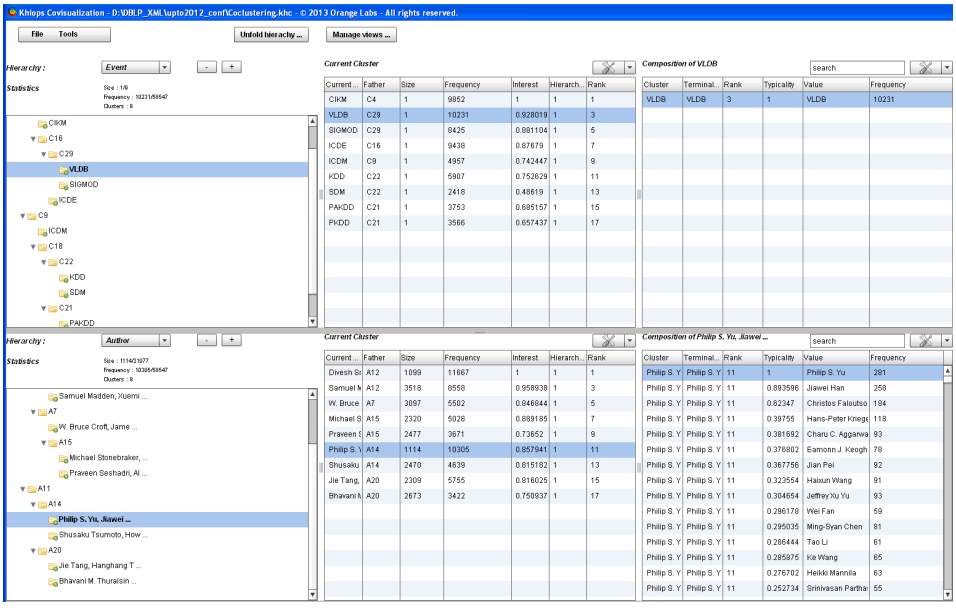

FIG. 1 – *Panneau principal de Khiops CoViz : (de gauche à droite), les hiérarchies des parties de deux dimensions (Author and Event), la liste des parties terminales des hiérarchies et la composition des parties sélectionnées.*

Choisir la granularité voulue pour visualiser la grille se fait au moyen de la fonctionnalité “unfold hierarchy” (voir figure 2 à droite). Un analyste peut contrôler le nombre de parties par dimension ou le taux d’information (comme décrit plus haut) par fusion optimale ou par fusion personnalisée (non-optimale). Selon les données d’applications, la grille optimale peut être composée de beaucoup de clusters (peut-être trop pour l’analyste) et fusionner des clusters augmente mécaniquement le nombre de valeurs par cluster : pour faciliter l’analyse des clusters, l’outil propose deux mesures, l’*intérêt* d’un cluster et la *typicité* d’une valeur dans un cluster. Ces deux mesures sont dérivées du critère *cost* (Guigourès, 2013) et sont utiles pour ordonner les clusters ou les valeurs de clusters par intérêt et se focaliser sur les composantes les plus intéressantes (voir figure 1).

Panneau de visualisation. Pour deux variables partitionnées X et Y à visualiser, les visualisations classiques, telles les “heat map” sont disponibles : il est ainsi possible de visualiser la fréquence des cellules, la probabilité jointe, la probabilité conditionnelle, la densité jointe ou conditionnelle. De plus, l’outil propose deux critères (dérivés de l’information mutuelle $MI(X, Y)$, voir Guigourès (2013) pour les définitions complètes) qui fournissent des informations supplémentaires sur les interactions entre variables :

- Contribution à l’information mutuelle (CMI) : CMI indique comment les cellules contribuent à l’information mutuelle $MI(X, Y)$; positivement (rouge), négativement (bleu), nullement (blanc) indique respectivement un excès d’interactions, un déficit ou aucune interaction particulière comparée à ce qui est attendu en cas d’indépendance des variables partitionnées.
- Contraste : Le contraste est dédié aux grilles 3D (ou plus). Étant données deux variables partitionnées fixées X, Y à visualiser et une partie P_i d’une troisième variable K, appelée contexte, le contraste met en lumière les cellules qui caractérisent P_i par rapport à toute les données. Comme précédemment, un contraste positif, négatif ou nul est différencié par les couleurs rouge, bleue et blanche.

FIG. 2 – *(Gauche) : Visualisation de la contribution à l’information mutuelle pour* $Year \times Event$*. (Droite) : Panneau de sélection de la granularité de la grille.*

3 Usage multiple des modèles en grille

L'analyse exploratoire des interactions entre variables par le biais de la visualisation des modèles en grilles est adaptée à plusieurs types de données et domaines d'applications (Bondu et al., 2013). Nous présentons une liste non-exhaustive des données d'applications typiques qui ont déjà été étudié via Khiops CoViz :

- Marketing : Les clients avec la liste des produits achetés ($customer \times product$)
- Web Mining : analyse de logs web pour identifier des comportements de navigation ($cookies \times webpages$)
- Télécommunications : Dimensionnement de réseau mobile par l'analyse des compte-rendus d'appels (CDRs) ($sourceAntenna \times targetAntenna$), e.g., analyse exploratoire des CDRs à l'échelle d'un pays (Guigourès, 2013).
- Fouille de textes : (co)clustering de textes ($Texts \times Words$)
- Fouille de graphes : Données multigraphes temporels ($SourceNodes \times TargetNodes \times Time$), e.g., analyse des locations de vélos à Londres (Guigourès et al., 2012).
- Clustering de données fonctionnelles (séries temporelles numériques ou catégorielles) : $TimeSeriesId \times Time \times Value$ ou $TimeSeriesId \times Time \times Event$, e.g., clustering de courbes (Boullé, 2012) ou analyse de consommation électrique (Boullé et al., 2012).

4 Conclusion & Travaux futurs

Nous présentons Khiops CoViz, un outil basé sur les modèles en grilles, pour identifier et visualiser les interactions intéressantes entre variables catégorielles et/ou numériques. Cependant, pour de nombreuses applications, l'analyste a souvent besoin de plus qu'une représentation matricielle des résultats, e.g., de représentations graphiques pour les données de graphes, des projections sur cartes pour les données géographiques : des extensions de l'outil par des plug-ins dédiés aux applications sont actuellement étudiées.

Références

Bondu, A., M. Boullé, et D. Gay (2013). Les modèles en grilles. Principes, évaluation, algorithmes et applications. *Tutorial given at EGC*.

Boullé, M. Data grid models for preparation and modeling in supervised learning. In *Hands-On Pattern Recognition : Challenges in Machine Learning, volume 1*.

Boullé, M. (2012). Functional data clustering via piecewise constant nonparametric density estimation. *Pattern Recognition 45*(12), 4389–4401.

Boullé, M., R. Guigourès, et F. Rossi (2012). Nonparametric hierarchical clustering of functional data. In *EGC (best of volume)*, pp. 15–35.

Dhillon, I. S., S. Mallela, et D. S. Modha (2003). Information-theoretic co-clustering. In *KDD'03*, pp. 89–98. ACM Press.

Guigourès, R. (2013). *Utilisation des modèles de co-clustering pour l'analyse exploratoire des données*. Ph. D. thesis, Université Paris 1 Panthéon-Sorbonne.

Guigourès, R., M. Boullé, et F. Rossi (2012). A triclustering approach for time evolving graphs. In *ICDM Workshops*, pp. 115–122.

Kolda, T. G. et J. Sun (2008). Scalable tensor decompositions for multi-aspect data mining. In *IEEE ICDM'08*, pp. 363–372.