Prototype de clustering exploratoire pour l'aide à la segmentation des clients

Adnan El Moussawi*,**, Philippe De Guis**, Arnaud Giacometti*,
Nicolas Labroche*, Arnaud Soulet*

*Université François Rabelais de Tours - {prénom}.{nom}@univ-tours.fr
**Group KALIDEA - {aelmoussawi, pdeguis}@kalidea.com

Résumé. Le clustering est une technique largement répandue pour la définition de profils dans le cadre de l'aide à la gestion de la relation client (CRM). Cependant, les outils classiques sont généralement limités, car ils ne prennent pas en compte la connaissance métier de l'analyste et ne permettent pas l'exploration interactive des données. Nous décrivons ici un prototype qui permet à un expert marketing d'explorer interactivement les données pour la recherche de profils des clients, mais aussi d'analyser les profils construits à l'aide de différentes visualisations synthétiques et d'étudier leurs évolutions au cours du temps.

1 Introduction

Nous considérons le problème de l'exploration de données par un expert, notamment dans le domaine du CRM où il cherche à explorer les données des clients pour construire des profils avec une sémantique compréhensible, pour leurs proposer des programmes spécifiques de fidélisation. Traditionnellement, deux méthodes peuvent être utilisées pour l'exploration de ces données : l'OLAP et les algorithmes de clustering.

Les outils OLAP permettent une analyse interactive des données avec des opérateurs de base pour sélectionner un sous-ensemble des données et spécifier les dimensions d'analyse pertinentes. Mais, ces outils ne permettent pas de révéler des modèles intéressants cachés dans les ensembles de données tels que des groupes de clients de comportements similaires.

Les algorithmes de clustering révèlent la structure naturelle des ensembles de données et peuvent résumer l'information en cas de données volumineuses. En revanche, ils manquent de l'interactivité qui permet de spécifier dynamiquement les sous-ensembles de données ou les attributs d'analyse qui intéressent l'expert. Dans les approches de clustering interactif telles que Balcan et Blum (2008) et Awasthi et al. (2013), l'interaction de l'utilisateur avec la méthode de clustering est par exemple limitée par des requêtes de scission et/ou fusion des clusters. Les approches du clustering, dites « semi-supervisée », permettent également l'interaction de l'expert via des contraintes au niveau des objets de données ou sur les caractéristiques de clusters (Grossi et al., 2016). Des travaux récents en collaboration avec l'entreprise KALIDEA ont contribué à proposer une nouvelle méthode de clustering semi-supervisée qui permet à l'expert de définir des préférences sur les attributs d'analyse (El Moussawi et al., 2016).

Le présent prototype a pour objectif de proposer un outil de segmentation semi-automatique avec une IHM adaptée à un expert non spécialiste de la fouille de données, simple à manipu-

ler, enrichi par des opérations d'exploration en bénéficiant des avantages de l'OLAP et du clustering et par des visualisations synthétiques pour faciliter l'interprétation des résultats.

2 Présentation du prototype

Architecture Ce prototype fait partie des travaux en R&D de l'entreprise KALIDEA. Il est constitué principalement d'une application Web permettant le paramétrage d'un cas d'analyse, la construction des données d'analyse, ainsi que l'exploration et l'analyse interactive des données et résultats. Cette interface est alimentée par deux autres modules, un module pour la gestion de données et un module pour la segmentation ou clustering, chacun d'eux comprenant un ensemble spécifique de fonctionnalités d'exploration.

Le module de gestion des données permet le stockage des données de base et la construction des données à analyser. Les données initiales des clients sont stockées dans un cube de données exploitable par un moteur d'OLAP, ce qui facilite pour l'expert la navigation dans les données et la construction des sous-ensembles de données à analyser.

Le module de clustering intègre deux algorithmes de clustering : l'algorithme K-Means traditionnel et l'algorithme MAPK-Means proposé par El Moussawi et al. (2016). Ce dernier algorithme permet la prise en compte de préférences d'expert sur les attributs d'analyse.

Fonctionnalités Le prototype doit être facile d'usage pour un expert métier non spécialiste de fouille de données et permettre une analyse exploratoire interactive des données et résultats. Pour ce faire, la version actuelle de notre outil intègre les fonctionnalités suivantes :

- Construction du jeu de données : elle s'effectue par exploration du cube OLAP contenant les données à analyser. L'expert peut spécifier son objet d'étude, les attributs d'analyse et la période d'analyse souhaitée.
- Fouille de données : la fouille permet la construction d'une segmentation, avec ou sans prise en compte des préférences d'un expert sur les attributs d'analyse.
- Analyse des résultats du clustering : l'analyse des résultats se fait en utilisant différents types de graphiques comme des histogrammes pour représenter le contenu des clusters, des camemberts pour évaluer la qualité des clusters construits, etc.
- Extraction des règles : l'expert peut faire appel à une méthode de construction d'arbre de décision pour expliciter la sémantique des clusters construits.
- Clustering exploratoire : l'expert peut interagir avec les résultats pour exclure un cluster de l'étude ou segmenter un cluster particulier.
- Représentation de l'évolution des clusters : elle permet à l'utilisateur d'effectuer des clusterings sur plusieurs pas de temps afin d'analyser l'évolution des clusters et des objets dans le temps.

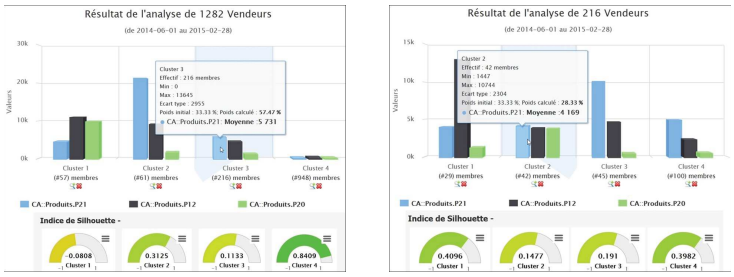
3 Cas d'usage

Nous nous sommes limités dans cette partie à un cas d'utilisation en mettant en avant les principales fonctionnalités du prototype. La base de données multidimensionnelles utilisée correspond à des données de ventes anonymisées d'un client de KALIDEA.

Construction du jeu données pour le clustering Après la sélection des données à étudier, l'utilisateur doit construire les données à segmenter. La construction commence par la sélection de l'objet d'étude à utiliser pour créer les clusters, qui correspond à une dimension dans le

cube. Pour notre exemple il s'agit des vendeurs. Ensuite l'utilisateur doit poursuivre en choisissant les attributs d'analyse, où un attribut correspond à une mesure quantitative associée à un membre d'une dimension : nous avons choisi le chiffre d'affaire (CA) réalisé sur les trois produits les plus vendus de la dimension 'Produits'. Pour finir, l'utilisateur peut définir une période d'analyse entre deux dates différentes. Les données résultantes de cette interaction avec le cube constituent le sous-ensemble de données en entrée de la méthode de clustering choisie.

Analyse des résultats du clustering Dans la figure 1, nous visualisons les résultats de l'exécution de la méthode MAPK-Means sur l'ensemble des données d'étude choisies précédemment (CA réalisé par 1282 Vendeurs sur les top-3 produits vendus P21, P12, P20) avec des poids uniformes (1a), puis nous avons appliqué un zoom sur le troisième cluster obtenu (1b).



(a) Clustering sur tous les vendeurs (b) Zoom sur le cluster 3

FIG. 1: Résultats du clustering avec MAPK-Means

Nous représentons chaque cluster par un histogramme formé de trois barres de couleurs différentes associées aux attributs d'analyse choisis (1a), la hauteur de chaque barre correspondant à la valeur moyenne d'un attribut. Nous pouvons ainsi résumer un cluster résultant. Par exemple le cluster 1 regroupe les vendeurs (5.07%) caractérisés par un CA moyen sur le produit P21 de 5k€, tandis que les CA réalisés sur les deux autres produits sont deux fois plus élevés. Cette visualisation permet aussi de voir les poids calculés par MAPK-Means pour tous les attributs d'analyse, comme le poids calculé pour le CA réalisé sur le produit P21 qui est égale à 0.56, ce qui indique il est plus discriminant que les autres attributs.
En dessous des histogrammes, une visualisation sous forme de camemberts permet de voir la qualité des clusters obtenus mesurée avec l'indice de silhouette. C'est un indicateur qui permet de guider l'exploration, où une mauvaise valeur indique qu'il est important de zoomer sur le cluster concerné. Une opération d'exploration de type « zooming » correspond à re-segmenter les individus du cluster. La figure (1b) correspond aux résultats d'un zoom sur le cluster 3 du clustering initial. Les « sous-clusters » obtenus ont des profils différents, ce qui explique la faible qualité du cluster 3 de (1a).

La construction d'un arbre de décision (Fig. 2) sur les résultats du clustering (1a), permet à l'expert de mieux appréhender la sémantique des clusters générés. Par exemple, l'arbre construit indique qu'un vendeur qui réalise un CA sur le produit P21 inférieur à 14378€ et un CA sur P12 inférieur à 17.95€ appartient au cluster 4.

Représentation d'évolutions des clusters Le diagramme de flux (Fig. 3) montre finalement les évolutions des clusters de vendeurs entre juin, juillet et août. Par exemple, 92% des vendeurs

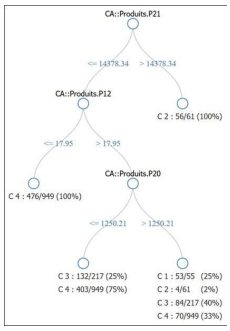


FIG. 2: Arbre de décision

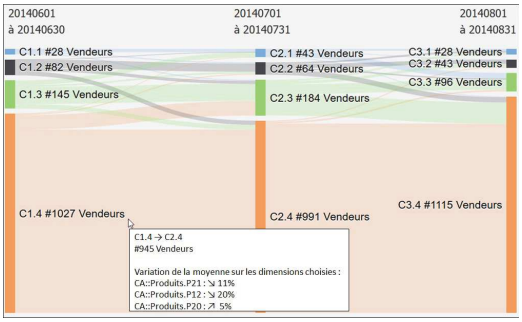


FIG. 3: Évolution des clusters entre les mois de juin, juillet et août

dans le cluster 4 en juin sont restés dans le même cluster en juillet, et nous pouvons remarquer que leurs CA moyens sur P21 et P12 ont diminué, tandis que le CA sur P20 a augmenté.

4 Conclusion

Nous avons développé un prototype de clustering exploratoire pour la segmentation des clients, adapté aux experts marketing en leurs permettant divers fonctionnalités d'exploration et d'analyse grâce aux techniques de fouille de données, d'OLAP et de visualisations. En perspective, nous souhaitons enrichir cet outil par des nouvelles fonctions analytiques ainsi que par des fonctions de recommandations de paramétrage, pour améliorer la qualité d'analyse.

Remerciement : Nous remercions l'ANRT pour leur soutien financier dans le cadre d'une thèse CIFRE (2014/0658).

Références

Awasthi, P., M. Balcan, et K. Voevodski (2013). Local algorithms for interactive clustering. *CoRR abs/1312.6724*.

Balcan, M.-F. et A. Blum (2008). Clustering with interactive feedback. In *Proceedings of the 19th ALT*, ALT '08, pp. 316–328.

El Moussawi, A., A. Cheriat, A. Giacometti, N. Labroche, et A. Soulet (2016). Clustering par apprentissage de distance guidé par des préférences sur les attributs. In *EGC'2016*, Volume 30, Reims, France, pp. 333–344.

Grossi, V., A. Romei, et F. Turini (2016). Survey on using constraints in data mining. *Data Mining and Knowledge Discovery*, 1–41.

Summary

Clustering is a widely used technique for customers segmentation in the CRM domain. However, actual tools are limited because they do not take into account experts knowledge and lack of an interactive exploration. We present a prototype that allows a marketing expert to interactively explore data and analytical attributes in the search of customer profiles. Moreover, our tool allows to study the evolution of profiles over time, and provides various synthetic visualizations that support the analysis task.