Vers l'intégration de la prédiction dans les cubes OLAP

Anouck Bodin-Niemczuk*, Riadh Ben Messaoud*
Sabine Loudcher Rabaséda**, Omar Boussaid**

Laboratoire ERIC, Université Lumière Lyon 2
5 avenue Pierre Mendès-France, 69676 Bron Cedex
*{abodin | rbenmessaoud}@eric.univ-lyon2.fr
**{ sabine.loudcher | omar.boussaid}@univ-lyon2.fr

L'analyse en ligne OLAP (*On Line Analytical Processing*) soutient les entrepôts de données dans le processus d'aide à la décision. Cependant, il n'existe pas d'outils pour guider l'utilisateur dans l'exploration, ni pour approfondir l'analyse vers l'explication et la prédiction.
Dans un processus décisionnel, un utilisateur peut vouloir anticiper la réalisation d'évènements futurs. Le couplage de la fouille de données avec la technologie OLAP permet d'assister l'utilisateur dans cette tâche pour l'extraction de nouvelles connaissances.
Nous discernons une dichotomie entre les travaux étudiés pour la prédiction dans l'OLAP. D'un côté, Chen et al. (2006) intègrent un processus complet de fouille de données pour l'élaboration d'un modèle de prédiction. D'un autre côté, Sarawagi et al. (1998) intègrent parfaitement le modèle dans l'environnement OLAP. La combinaison des deux approches permettrait une réelle intégration de la prédiction à l'analyse en ligne.
Nous proposons un cadre de prédiction OLAP fondé à la fois sur la philosophie OLAP et sur la fouille de données. Via une technique de type arbre de régression, l'utilisateur peut prédire la valeur de la mesure d'un nouveau fait selon un contexte d'analyse défini par ses soins. Nous nous plaçons dans le cadre du *"What if analysis"* où le procédé de projection dans l'avenir illustre une démarche centrée sur l'utilisateur OLAP. Nous utilisons un processus complet d'apprentissage automatique et exploitons les résultats obtenus dans le cube de données OLAP.

Nous réalisons un premier pas vers un cadre de prédiction OLAP en y associant les arbres de régression. Notre démarche se résume de la manière suivante :
Le point de départ est un contexte d'analyse $\mathcal{C}'$ (sous-cube) avec n faits OLAP observés selon la mesure quantitative M_q, défini par l'utilisateur au sein d'un cube de données $\mathcal{C}$. Pour la construction et la validation du modèle, le contexte d'analyse est segmenté en deux : 70% des faits servent à l'apprentissage et 30% à l'évaluation du modèle. Les critères d'évaluation sont le taux d'erreur moyen et la réduction de l'erreur.
Soit $R(X \Rightarrow Y; S; \sigma)$ une règle de décision obtenue dans le modèle. X est une conjonction et/ou disjonction de modalités. Y est la valeur moyenne prédite pour la mesure M_q sachant X. S est le support de la règle et σ est l'écart-type de M_q dans l'ensemble d'apprentissage vérifiant X. Pour exploiter les règles dans l'environnement OLAP nous procédons ainsi : pour intégrer la règle $R(X \Rightarrow Y, S, \sigma)$ dans le sous-cube $\mathcal{C}'$, on affecte à la cellule c vide qui vérifie X, la valeur prédite Y. Les agrégats à un niveau hiérarchique supérieur peuvent alors êtres calculés en y intégrant les valeurs prédites aux niveaux inférieurs. Afin de valoriser le