Une approche connexionniste pour l'extension de l'OLAP à des capacités de prédiction

Wiem Abdelbaki***, Sadok Ben Yahia*, Riadh Ben Messaoud**

*Faculté des Sciences de Tunis, Tunisie.
**Faculté des Sciences Économiques et de Gestion de Nabeul, Tunisie.

Résumé. Les outils de l'analyse en ligne (OLAP) permettent à l'utilisateur de réaliser des tâches exploratoires dans les cubes de données. Cependant, ils n'offrent aucun moyen pour la prédiction ou l'explication des faits. En vue de renforcer le processus de l'aide à la décision, plusieurs travaux ont proposé l'extension de l'analyse en ligne à des capacités plus avancées. Dans cet article, nous proposons une nouvelle approche d'extension de l'analyse en ligne à des capacités de prédiction à deux phases. La première est une phase de réduction des dimensions des cubes de données, qui repose sur l'analyse en composantes principales (ACP). La deuxième est une phase de prédiction dans laquelle nous introduisons une nouvelle architecture de percéptrons multicouches (PMC). Notre étude expérimentale a montré une capacité de prédiction prometteuse, ainsi qu'une bonne robustesse dans le cas d'un cube épars.

1 Introduction

Les entrepôts de données représentent une solution convenable au problème du stockage et de l'analyse des données historisées d'une entreprise (Inmon, 1996). En fait, ils permettent de stoker les données dans des structures multidimensionnelles appelées *cubes de données*. Ces cubes, contiennent des données, qui sont déjà consolidées et intégrées (Inmon, 1996). Par ailleurs, la visualisation d'un cube dévoile généralement une structure éparse. En effet, les cellules vides reflètent des événements passés, qui n'ont pas eu lieu à l'instant présent ou des événements futurs, qui n'ont pas encore eu lieu.

Nous pensons qu'il serait très intéressant pour l'utilisateur de pouvoir anticiper les valeurs de la mesure d'une cellule vide pour une meilleure prise de décision. Par exemple, il serait très utile à une société de vente de voitures de prédire le chiffre d'affaire, que peut réaliser une nouvelle agence dans une nouvelle ville avant de lancer ce projet. Cet indicateur encouragera, ou non, l'investissement de cette société dans cette agence. En effet, l'information manquante, qui concerne une cellule, est potentiellement existante dans son voisinage. Cependant, la technologie OLAP n'offre aucun moyen pour l'explication ou la prédiction des faits.

En réponse à ce constat, plusieurs études proposent d'étendre l'OLAP pour répondre à différents objectifs comme l'exploration des cubes de données (Sarawagi et al., 1998) et l'extraction des règles d'association (Ben Messaoud, 2006). Récemment, de nouvelles recherches proposent d'étendre OLAP à des capacités de prédiction afin d'anticiper des événement futurs (Bodin-Niemczuk et al., 2008; Chen et al., 2005).

Dans cet article, nous essayons d'intégrer les Perceptrons Multicouches (PMC) dans l'environnement OLAP pour l'étendre à des capacités de prédiction. Pour cela, nous proposons une nouvelle approche de prédiction dans les cubes OLAP, qui s'articule autour de deux phases. La première est une phase de réduction des dimensions, qui respecte la structure multidimensionnelle des cubes OLAP et en génère des ensembles d'apprentissage réduits, adaptés à leur contexte particulier, en exploitant l'analyse en composantes principales (ACP). Quant à la deuxième phase, elle cible la prédiction de la mesure des faits inexistants en exploitant une nouvelle architecture de Perceptrons Multicouches.

Le reste de l'article est organisé comme suit. Dans la section 2, nous présentons une étude de l'état de l'art. Dans la section 3, nous positionnons le contexte de notre travail et nous présentons une vue globale sur notre approche. Nous formalisons les étapes de notre proposition dans la section 4. Dans la section 5, nous présentons les résultats expérimentaux encourageants, obtenus à partir d'une base de données réelle de recensement. Enfin, nous dressons une conclusion et discutons les perspectives de notre approche dans la section 6.

2 État de l'art

Récemment, plusieurs recherches ont abordé l'extension de l'OLAP à des capacités plus sophistiquées. Dans ce qui suit, nous détaillons celles qui sont étroitement liées à la prédiction dans les cubes OLAP.

Sarawagi et al. (1998) proposent une approche basée sur la modélisation log-linéaire pour estimer la mesure en détectant ses valeurs exceptionnelles. Les auteurs optimisent leur proposition par la réutilisation des modèles de niveaux hiérarchiques les plus bas pour la construction de ceux, qui viennent de niveaux d'agrégation plus élevés. Palpanas et al. (2005) exploitent les faits agrégés pour prédire les faits détaillés. Ils produisent des estimations des résultats d'une requête en utilisant le principe de l'entropie maximale et un algorithme itératif d'ajustement proportionnel. Chen et al. (2005) introduisent des nouveaux cubes appelés *prediction cubes*. La mesure de ces derniers reflète une probabilité, un score ou une distribution de la mesure originale. Les auteurs ont réutilisé la proposition d'optimisation de Sarawagi et al. (1998). Chen et al. introduisent les cubes de régression dans le cadre de compression des cubes de données. Ils contiennent des mesures compressibles générées en se basant sur la régression linéaire. Ces nouvelles mesures reflètent les variations des données initiales. Bodin-Niemczuk et al. (2008) proposent une approche de prédiction dans les cubes OLAP, qui utilise les arbres de régression pour prédire les valeurs de la mesure des faits inexistants. Les auteurs ne traitent pas l'élagage de l'arbre pour l'optimisation de leur algorithme, qui doit faire face au passage à l'échelle des données volumineuses.

Le Tableau 1 résume les approches sus-citées. Nous notons par (+) si l'approche n'exige pas une hypothèse particulière, implémente une optimisation, exploite une réduction de données et/ou fournit un résultat explicite. Nous notons par (-) dans les situations opposées.

Nous remarquons que la plupart des approches, dont l'objectif principal est différent de la prédiction, ne fournissent pas de résultats de prédiction explicites. Cela reste toutefois justifiable, si la recherche remplit ses objectifs principaux. D'autre part, nous constatons que la volumétrie de données est le défi principal dans toutes les approches, qui ont affaires à des cubes de données. La plupart des approches envisagent une de deux solutions. La première consiste en l'intégration d'un algorithme d'optimisation au sein de l'algorithme principal. Quant à la deuxième, elle repose sur la réduction des données. Cependant, nous constatons que la propo-

sition de Bodin-Niemczuk et al. (2008) est la seule approche, qui n'adapte aucune réduction ni optimisation. En conséquent, le rendement de cette approche peut se dégrader face au cubes de données du monde réel caractérisés par une volumétrie importante.

	Objectif	Hypothèse	Optimisation	Réduction	Résultat
Sarawagi et al. (1998)	Exploration	-	+	-	-
Palpanas et al. (2005)	Compression	+	-	+	+
Chen et al. (2005)	Prédiction	+	+	-	-
Chen et al.	Compression	-	-	+	-
Bodin-Niemczuk et al. (2008)	Prédiction	+	-	-	+
Notre approche	Prédiction	+	-	+	+

TAB. 1: Synthèse des propositions de prédiction dans les cubes de données OLAP

Notre proposition se place dans le cadre du couplage entre l'OLAP et la fouille de données. Elle n'exige pas d'hypothèse particulière sur les cubes traités. De plus, elle exploite un algorithme de réduction des dimensions vers le contexte des cubes de données OLAP. D'autre part, notre approche fournit des valeurs explicites de la mesure des faits inexistant, supportée par des indicateurs de qualité.

3 Objectifs et vue globales

Notre proposition consiste à intégrer un modèle de prédiction dans l'environnement OLAP afin d'enrichir le processus de prise de décision. Nous cherchons à prédire des mesures inexistantes tout en traitant des cubes de données volumineux. Les objectifs principaux de notre approche peuvent se résumer dans les points suivants :

1. Générer des ensembles d'apprentissage réduits à partir du cube de données initial ;
2. Adapter les PMCs à la structure multidimensionnelle des données ;
3. Fournir des valeurs explicites aux mesures ciblées ;
4. Accompagner les mesures prédites avec des indicateurs de qualité.

Le fait que les PMCs ne sont pas conçus pour les structures multidimensionnelles de données, a empêché leur exploitation dans le cadre des cubes OLAP même s'ils ont prouvé leur efficacité avec les données bidimensionnelles (Sharda, 1994; Haykin, 1999). De plus la combinaison de la structure multidimensionnelle et la volumetrie importante engendre des mesures fortement corrélées, bruitées et qui contiennent trop de redondance. Selon Bishop (1995), ces caractéristiques peuvent dégrader la capacité de généralisation des PMCs.

En réponse à ce constat, la première étape de notre approche consiste à générer des ensembles d'apprentissage réduits à partir du cube initial. Pour ce faire, nous exploitons l'Analyse en Composantes Principales (ACP), qui est une technique descriptive (Hotelling, 1933). Elle calcule des combinaisons linéaires des variables d'une matrice pour en générer des nouveaux axes factoriels, qui contiennent la plus grande partie de la variabilité, ce qui permet de se retrouver avec des données décorrélées, appelées *composantes principales*. Récemment, de plus en plus de travaux exploitent cette approche factorielle pour des objectifs de prédiction, en considérant les composantes principales comme des variables indépendantes des systèmes de prédiction (Tshilidzi, 2009; Wang et al., 2009). Nous suivons cette piste comme une étape de pré-traitement, qui assure la génération des ensembles d'apprentissage complémentaires, qui préservent la variation de la mesure et la sémantique reliant les différentes modalités.

Afin de faire face à la structure multidimensionnelle de données, nous proposons une nouvelle architecture de PMC basée sur une interconnexion de sous-réseaux. Elle permet d'impliquer des ensemble disjoints de données dans le même processus d'apprentissage sans avoir recours à les fusionner et pourtant en générer une unique valeur pour chaque mesure cibleé.

Il est important de noter que notre approche n'est pas une technique de complétion. Elle ne cible pas le remplissage de toutes les cellules vides du cube. L'objectif principal de notre proposition est de fournir une valeur à n'importe quelle mesure requise par l'utilisateur afin d'anticiper un indicateur décisif comme un chiffre d'affaires futur, comme dans l'exemple décrit précédemment dans la section 1.

4 Formalisation de notre approche

4.1 Notations générales

Dans le reste de cet article, nous utilisons les notations relatives à la structure d'un cube de données de Ben Messaoud (2006). Nous considérons alors, que $\mathcal{C}$ est un cube de données caractérisé par les propriétés suivantes :

- $\mathcal{C}$ est constitué d'un ensemble non vide de d dimensions $\mathcal{D} = \{D_i\}_{(1 \leq i \leq d)}$;
- $\mathcal{C}$ contient un ensemble non vide de m mesures $\mathcal{M} = \{M_q\}_{(1 \leq q \leq m)}$;
- Chaque dimension D_i contient l_i modalités catégorielles. i.e., $\mathcal{A} = \{a_1^i, ..., a_t^i, ..., a_{l_i}^i\}$ est l'ensemble des modalités de la dimension D_i.

4.2 Réduction des dimensions d'un cube de données

Pour préserver les avantages de l'ACP, nous essayons de profiter de l'aspect multidimensionnel des cubes de données OLAP pour en extraire des ensembles de données, exploitables dans le cadre de l'ACP classique. Pour ce faire, nous proposons d'aplatir le cube de données en un ensemble de tables bidimensionnelles, qui préserve les variations de la mesure et la sémantique qui relie les modalités d'une même dimension.

L'aplatissement d'un cube de données revient à appliquer l'opérateur OLAP *Slice* sur le cube initial. Cependant, à partir d'une unique tranche OLAP nous obtenons deux matrices, dont l'une correspond à la transposée de l'autre. En appliquant une technique issue de l'analyse numérique comme l'ACP, ces deux matrices fournissent des résultats différents. Pour respecter l'idée fondamentale de notre approche, qui cible la couverture de toutes les variations de la mesure, nous traitons séparément toutes les matrices des chaque tranche OLAP. Pour ce faire, nous proposons la notion du *cube-face* pour capter toutes les configurations possibles d'un cube et la notion de *ACP-slice* pour capter les variations de la mesure selon les modalités de ses différentes dimensions.

Définition 1 (Un Cube-face) $Cf(D_c, D_v, (D_{s_1}, \ldots, D_{s_i}, \ldots, D_{s_{d-2}}))$ d'un cube de données $\mathcal{C}$, est une vue spécifique de ce même cube. Il est identifiable par les positions géométriques de ses dimensions, que nous notons :

- La dimension *Clé* D_c.
- La dimension *Variante* D_v.
- L'ensemble des dimensions *Slicers* D_s, qui est composé par $(d-2)$ dimensions notées D_{s_i} ; avec $i \in [1, d-2]$ et tous les D_{s_i} doivent être strictement différents de D_c et D_v.

Définition 2 (ACP-slice) Un ACP-slice $S(D_c, D_v, (a_{t_1}^o, \ldots, a_{t_i}^p, \ldots, a_{t_{d-2}}^q))$ est la tranche résultante de l'opérateur *Slice* sur le cube-face $Cf(D_c, D_v, (D_{s_1}, \ldots, D_{s_i}, \ldots, D_{s_{d-2}})$, tel que ; $a_{t_1}^o a_{t_i}^p et a_{t_{d-2}}^q$ appartiennent respectivement à $D_{s_1}, D_{s_i} et D_{s_{d-2}}$,

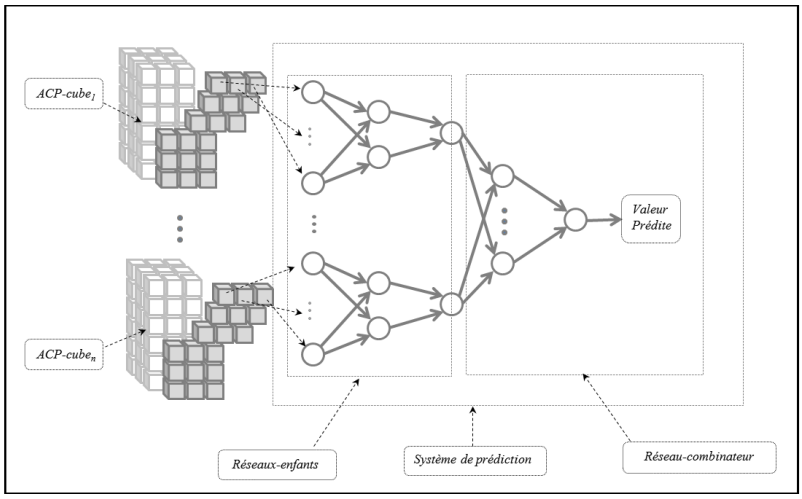

FIG. 1: Notre système de prédiction

Pour la réduction des dimensions du cube, nous commençons par fixer tous les cube-faces d'un cube de données. Ensuite, nous procédons à l'extraction de tous les ACP-slices à partir de ces cube-faces. Enfin, nous appliquons l'ACP sur chaque ACP-slice d'une manière itérative.

Dans notre situation, la réduction du cube est une phase de pré-traitement dont l'objectif est de préparer les données pour une phase de prédiction. Par conséquent, un niveau minimum d'homogénéité entre les deux phases est strictement requis, car le résultat de l'une fera l'objet de l'entrée de l'autre.

Par la suite, une contrainte sur le nombre de composantes principales à garder s'avère nécessaire. Il faut donc garder le même nombre de composantes pour tous les ACP-slices appartenant au même cube-face. Ceci va générer des ensembles de tables de coordonnées factorielles de mêmes dimensions correspondants chacune à un ACP-slice particulier. Schématiquement, ils forment des cubes adjacents au cube-faces. Nous les appelons alors *ACP-cubes*.

Notons que les cellules, qui appartiennent à la même ACP-slice et partagent la même modalité de la dimension *variante* ; partagent les mêmes coordonnées factorielles, ce qui met en évidence l'aspect de la réduction.

4.3 Prédiction de la mesure

Dans notre situation, les faits contenus dans un ACP-slice considèrent les données contenues dans leur ACP-cube comme prédicateurs. Pour l'apprentissage d'un système de prédiction, une approche naïve consiste à fusionner tous ces ensembles d'apprentissage dans une seule grande base. Cependant, cette approche va engendrer la perte de l'unicité des informations apportées par chaque ACP-cube.

Pour cela, nous proposons une nouvelle architecture de réseaux des neurones, qui se base sur une interconnexion d'un ensemble de sous-réseaux, que nous appelons réseau-enfants. Chaque réseau-enfant possède une seule neurone de sortie et s'alimente à partir d'un unique ACP-cube. Tous les neurones de sortie des réseaux-enfants sont rassemblés par seul réseau, que nous appelons réseau-combinateur. Il les considère comme ses entrées, comme le montre la Figure 1. Ainsi, nous garantissons que chaque ACP-cube contribue d'une manière unique dans la

prédiction en alimentant un réseau distinct. Le réseau-combinateur réunit les valeurs de sorties de tous les réseau-enfants, pour en générer une seule valeur de conclusion, qui représente la valeur prédite de la mesure.

Entrée: Cube de données C, ACP-cubes de C pcc, Fait à prédire f, Faits d'apprentissage $a[]$, Nombre de ACP-cubes n;
Sortie: Système de prédiction ajusté;
1 : **pour** i **de** 1 **à** n
2 : Rne_i←**nitialiser_réseau**(**Taille**(pcc_i)); //Initialisation des réseaux-enfants
3 : **Fin pour**
4 : Rnc←**Initialiser**(n); //Initialisation du réseau-combinateur
5 : **pour** i **de** 1 **à n**
6: **tant qu**'il existe d'instances d'apprentissage
7 : **Get_input**(pcc_i); //Extraction des coordonnées factorielles à partir de l'ACP_cube
8 : **Get_output**(a_i); //Extraction de la mesure du fait courant
9 : **propager**(Rne_i, $input_i$); //Propagation des coordonnées factorielles dans les Rne_i
10: **rétropropager**(Rnc); //Ajustement des poids de Rnc
11: **Fin tant que**
12: **Fin pour**
13: **tant qu**'il existe d'instances d'apprentissage
14: **pour** i **de** 1 **à** n
15: **propager**(Rne_i, **propager**(Rne_i, $input_i$)); //Propagation des coordonnées factorielles dans les Rne_i
16: **rétropropager**(Rne_i); // Ajustement des poids de Rne_i
17: **Fin pour**
18: **Fin tant que**

FIG. 2: Algorithme d'apprentissage de notre système

Notons $R(D_c, D_v, (D_{s_1}, \ldots, D_{s_i}, \ldots, D_{s_{d-2}}))$ le réseau-enfant, qui correspond au cube-face $Cf(D_c, D_v, (D_{s_1}, \ldots, D_{s_i}, \ldots, D_{s_{d-2}}))$.

Nous limitons les structures unitaires de nos réseaux à trois couches (dont une cachée) pour tous les sous-réseaux, qui forment notre système de prédiction. En effet, plusieurs études théoriques et empiriques comme celle de Hornik et al. (1989), ont montré qu'une seule couche cachée est suffisante pour parvenir à une approximation satisfaisante de n'importe quelle fonction non linéaire. Pour l'apprentissage de notre système, nous utilisons une version adaptée de l'algorithme de rétropropagation du gradient proposé par Rumelhart et Mcclelland (1986). D'après Haykin (1999), cet algorithme a fait ses preuves dans le cadre de plusieurs applications. Nous l'accompagnons avec la méthode du gradient conjugué comme méthode d'apprentissage et de la fonction d'activation sigmoïde comme fonction d'activation.

L'algorithme d'apprentissage de notre système est présenté dans la figure 2. Dans cet algorithme l'apprentissage est réparti sur deux étapes. La première est l'apprentissage des réseau-enfants. Dans cette étape, chaque réseau-enfant reçoit les coordonnées factorielles extraites depuis l'ACP-cube, qui lui est associé, comme valeurs d'entrée. Une fois que tous les réseau-enfants sont ajustés, ils sont alimentés d'une manière simultanée, depuis leurs ACP-cubes respectifs. Leurs valeurs de sorties sont injectées dans le réseau-combinateur, qui ajuste ses poids synaptiques selon l'algorithme de rétropropagation du gradient. Ce processus se déroule d'une manière itérative jusqu'à la convergence de l'apprentissage. Quant à la prédiction, elle se déroule d'une manière similaire à l'apprentissage du réseau-combinateur sauf, que la rétropropagation ne se déroulera plus.

La contribution principale de cette nouvelle architecture est qu'elle implique plusieurs ensembles de données dans un même processus d'apprentissage, sans les fusionner. Ce fait renforce l'unicité de la contribution de chaque ACP-cube et souligne la capacité de cette architecture de traiter les structures multidimensionnelles.

5 Étude expérimentale

Afin de valider notre approche, nous avons implémenté un prototype expérimental de notre système. Nous avons exploité la base de données American Community Surveys 2000-2003 *, qui est une base de données réelle relative au recensement américain, tout en l'adaptant au contexte des entrepôts de données.

Nous avons considéré un cube de données, qui contient 3.8 millions faits et 3 dimensions *Location*, *Origin* et *Education*. La dimension *Location* contient les données géographiques du recensement. La dimension *Origin* concerne la structure raciale de la population américaine. Quant à la dimension *Education*, elle contient les informations, qui concernent les niveaux d'éducation. Nous étudions la mesure *person count* en fixant les niveaux hiérarchiques ; *State*, *Race* et *Education*. Ces niveaux contiennent, respectivement, 51, 10 et 14 modalités.

Nous commençons par l'application de notre approche de réduction, nous nous retrouvons avec 6 ACP-cubes, qui contiennent respectivement 10, 12, 4, 3, 4 et 3 composantes principales. Pour l'application de la phase de prédiction, nous utilisons la technique de validation croisée à 5-fold et la moyenne des RMSEs (racine carrée de la moyenne des carrés des erreurs) comme critère d'évaluation du système. Pour tous les réseau-enfants, ainsi que pour le réseau-combinateur, nous utilisons une architecture multicouches, qui contient une seule couche cachée, ainsi qu'un taux d'apprentissage de 0,05.

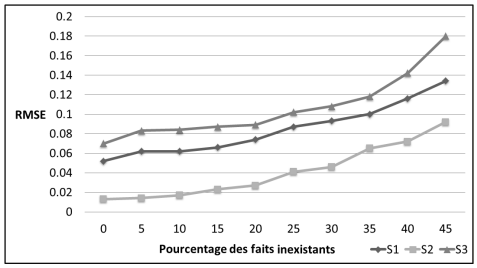

FIG. 3: Performances du système

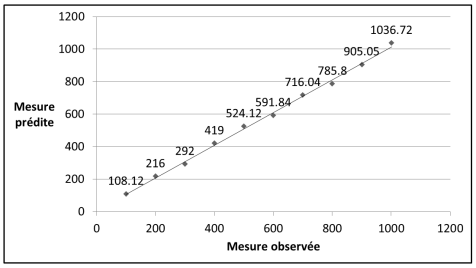

FIG. 4: Qualité de la prédiction

Notre étude expérimentale s'étale sur trois expériences, qui tentent de répondre à des objectifs différents. La première cible l'étude de la robustesse de notre système en se confrontant à différents taux de faits inexistants. Pour ce faire, nous concevons trois systèmes distincts, que nous notons $S1$, $S2$ et $S3$. Le nombre de neurones de la couche cachée de $S1$ est égal à un tier de celui de $S3$. Quant à celui de $S2$, il est égal à un demi de celui de $S3$. Nous évaluons les performances de ces systèmes en fonction de la moyenne des RMSEs des ensembles du test, tout en variant graduellement le pourcentage des faits inexistants dans la base d'apprentissage.

D'après la Figure 3, nous constatons que la RMSE suit la même tendance pour les 3 systèmes. De plus, $S2$ dépasse à la fois $S1$, qui contient moins de neurones dans ses couches

$*$. American Community Surveys est accessible depuis le site officiel IPUMS-USA (Integrated Public Use Microdata Series) ; http ://sda.berkeley.edu

cachées et $S3$, qui contient plus de neurones dans ses couches cachées. Ceci affirme que le choix du nombre de neurones dans les couches cachées doit être judicieusement fixé. Si le nombre est insuffisant, alors le modèle sera incapable de modéliser les données et la qualité du résultat sera modeste. Inversement, un problème de surapprentissage peut apparaître si les couches cachées contiennent plus qu'il faut de neurones

D'autre part, nous constatons que les valeurs de RMSE évoluent d'une manière assez monotone entre 0% et 20%. À partir de 20%, l'évolution de RMSE devient plus importante. Ceci s'explique par l'existence d'un nombre suffisant d'instances valides pour soutenir le processus d'apprentissage dans le premier intervalle. À partir d'un taux de faits inexistants de 20%, le système commence à rencontrer des difficultés dans la capture des motifs.

L'objectif de la deuxième expérience est l'étude de la qualité de la prédiction de notre système. Nous y exploitons $S2$, qui vient de prouver son efficacité dans l'expérience précédente. Pour cela, nous essayons de prédire les mesures des faits, qui n'ont participé ni au processus de réduction, ni à celui d'apprentissage. Pour bien présenter nos résultats, nous considérons des valeurs séparées par des intervalles réguliers. La courbe obtenue, présentée dans la Figure 4, montre que les valeurs prédites possèdent des distances minimales de la droite *Mesure observée=Mesure prédite*, ce qui confirme une bonne précision de prédiction.

Dans la troisième expérience nous essayons d'étudier l'apport de notre nouvelle architecture. Pour cela, nous étudions les valeurs des RMSEs obtenues de différents sous-réseaux, qui forment notre système. Nous exposons les résultats obtenus dans le Tableau 2.

	R(L,O,(E))	R(L,E,(O))	R(E,O,(L))	R(E,L,(O))	R(O,E,(L))	R(O,L,(E))	Rc	Rc_restreint
RMSE_Train	0.390	0.045	0.690	0.032	0.098	0.035	0.022	0.040
RMSE_Test	0.482	0.068	0.630	0.061	0.183	0.027	0.025	0.032

TAB. 2: Performances détaillées de notre nouvelle architecture

Nous trouvons que les valeurs des RMSEs varient d'une manière remarquable d'un réseau-enfant à un autre. Ceci est toutefois justifiable par la nature structurelle de nos données. Nous trouvons que les deux réseau-enfants, qui fournissent les RMSEs les plus importants, par conséquent les qualités de prédiction les moins bonnes, sont $R(E, O, (L))$ et $R(L, O, (E))$. Nous constatons, que ces deux réseau-enfants considèrent *Origin* comme leur dimension *variante*. En effet, dans notre cube la dimension *Origin* est la dimension la moins riche en nombre de modalités (10 modalités). Par la suite, les réseau-enfants qui la considèrent, en tant que dimension *variante* disposent du nombre d'instances d'apprentissage le plus réduit par rapport aux autres réseau-enfants. Inversement, nous trouvons que $R(E, L, (O))$ et $R(O, L, (E))$, qui considèrent la dimension *Location* en tant que dimension *variante*, produisent les valeurs de RMSE les plus réduites. Nous constatons, que plus le nombre des instances d'apprentissage est important, plus la qualité d'apprentissage est meilleure.

Ces résultats sont obtenus dans le cadre d'une étude détaillée de notre système. En effet, seules les résultats produits par le réseau-combinateur Rc reflètent la qualité de notre système globale. Or ce dernier surpasse tous les sous-réseaux dans la phase d'apprentissage, ainsi que dans la phase du test. Ainsi notre nouvelle architecture a permis de produire de meilleures qualités à partir des sous-réseaux de qualités différentes.

Suite à ces résultats, une question logique s'impose : Quelles seront les performances de notre système si nous éliminons les réseau-enfants, qui fournissent les résultats les moins bons de l'analyse. Pour répondre à cette question, nous avons entraîné le réseau $Rc\_restreint$,

dans lequel nous avons éliminé $R(E, O, (L))$ et $R(L, O, (E)$ de l'analyse. Nous avons trouvé, que la qualité de la prédiction se dégrade par rapport à Rc, comme il est indiqué dans le Tableau 2. Ceci s'explique par le fait que les informations non pertinentes fournies par les deux réseau-enfants éliminés, se transforment en informations importantes pour le système globale. Autrement, ensemble, les valeurs prédites à partir d'un cube-face, servent à raffiner la qualité du système globale même si elles ne sont pas exploitables individuellement.

6 Conclusion et perspectives

Dans cet article, nous avons proposé une nouvelle approche d'extension de l'OLAP à des capacités de prédiction en y intégrant un algorithme issu de l'apprentissage automatique. Notre contribution s'articule autour de deux axes. Le premier est la proposition d'une solution au problème de la dimensionnalité importante des cubes de données, qui handicape les algorithme de la fouille. Pour cela, nous avons introduit une nouvelle proposition de réduction des dimensions de cubes de données, qui permet de générer des prédicateurs représentatifs, en exploitant l'ACP. Quant au deuxième axe, il s'agit d'une nouvelle architecture de réseaux de neurones, qui permet de considérer des ensembles d'apprentissage disjoints, sans avoirs recours à les fusionner. Ceci a permis de préserver l'unicité de la contribution de chaque ensemble de prédicateurs dans le processus d'apprentissage. Ensuite, nous avons réalisé une étude expérimentale détaillée, qui s'étale sur trois expériences, de différents objectifs. Cette étude a montré les bonnes performances de prédiction. De plus, elle a permis d'expliquer les différents résultats obtenus à partir de différents sous-réseaux, qui composent notre système globale. D'un autre côté, cette étude a souligné la nécessité d'un bon calibrage du système de prédiction pour pouvoir fournir de bons résultats.

Ce travail ouvres les portes sur plusieurs perspectives. Tout d'abord, nous pensons que nous devons inclure une proposition d'explication des motifs d'apparition des faits inexistants dans les cubes de données, dans une approche similaire à celle proposée par (Ben Othman et Ben Yahia, 2008), qui est réalisée dans le cadre des données bidimensionnelles. Ensuite, nous pensons que nous devons impliquer la structure hiérarchique des cubes de données d'une manière plus profonde dans notre système. Ceci permettra d'exploiter les différents niveaux d'agrégation pour prédire les mesures des faits de niveaux inférieurs et/ou supérieurs. Enfin, nous pensons que la modélisation d'une relation théorique entre le système de réduction et celui de prédiction sera très utile, pour l'optimisation de notre système.

Remerciements

Ce travail a été partiellement financé par le projet CMCU 11G1417.

Références

Ben Messaoud, R. (2006). *Couplage de l'analyse en ligne et de la fouille de données pour l'exploration, l'agrégation et l'explication des données complexes*. Ph. D. thesis, Université Lumiére Lyon 2, Lyon, France.

Ben Othman, L. et S. Ben Yahia (2008). Yet Another Approach for Completing Missing Values. In S. Ben Yahia, E. Mephu Nguifo, et R. Belohlavek (Eds.), *In Proceeding of the International Conference on Concept Lattices and Their Applications*, Volume 4923 of *Lecture Notes in Computer Science*, pp. 155–169. Hammamet, Tunisia.

Bishop, C. (1995). *Neural Networks For Pattern Recognition*. Oxford University Press.

Bodin-Niemczuk, A., R. B. Messaoud, S. L. Rabaséda, et O. Boussaid (2008). Vers l'intégration de la prédiction dans les cubes OLAP. In *EGC*, pp. 203–204.

Chen, B.-C., L. Chen, Y. Lin, et R. Ramakrishnan (2005). Prediction Cubes. In *Proceedings of the 31st International Conference on Very large Data Bases*, VLDB '05, pp. 982–993.

Chen, Y., G. Dong, J. Han, J. Pei, B. W. Wah, et J. Wang. Regression Cubes with Lossless Compression and Aggregation. *IEEE Trans. on Knowl. and Data Eng. 18*.

Haykin, S. (1999). *Neural Networks : a Comprehensive Foundation*. Prentice Hall International Editions Series. Prentice Hall.

Hornik, K., M. Stinchcombe, et H. White (1989). Multilayer Feedforward Networks are Universal Approximators. *Neural Networks 2*(5), 359 – 366.

Hotelling, H. (1933). Analysis of a Complex of Statistical Variables into Principal Components. *Journal of Educational Psychology 24*(7), 498 – 520.

Inmon, W. H. (1996). *Building the Data Warehouse*. John Wiley & Sons.

Palpanas, T., N. Koudas, et A. Mendelzon (2005). Using Datacube Aggregates for Approximate Querying and Deviation Detection. *IEEE Trans. on Knowl. and Data Eng. 17*, 1465–1477.

Rumelhart, D. E. et J. L. Mcclelland (1986). *Parallel Distributed Processing : Explorations in the Microstructure of Cognition : Foundations (Parallel Distributed Processing)*.

Sarawagi, S., R. Agrawal, et N. Megiddo (1998). Discovery-driven Exploration of OLAP Data Cubes. In *Proceedings of the* 6^{th} *International Conference on Extending Database Technology (EDBT'1998)*, Valencia, Spain. Springer.

Sharda, R. (1994). Neural networks for the ms/or analyst : An application bibliograph. *Interfaces 24*(2), 116–130.

Tshilidzi, M. (2009). *Computational Intelligence for Missing Data Imputation, Estimation, and Management : Knowledge Optimization Techniques*. Hershey, PA : Information Science Reference - Imprint of : IGI Publishing.

Wang, Z., J. Xu, F. Lu, et Y. Zhang (2009). Using the Method Combining PCA with BP Neural Network to Predict Water Demand for Urban Development. In *Proceedings of the 2009 Fifth International Conference on Natural Computation - Volume 02*, Washington, DC, USA, pp. 621–625. IEEE Computer Society.

Summary

On-line analytical processing (OLAP), allows users to explore, structure and navigate into data cubes. However, OLAP is not capable to explain or predict the facts of a data cube. In order to enhance the decision-making process, many research proposes to extend OLAP to advanced capabilities. We introduce in this paper a novel approach attempting to extend OLAP to prediction capabilities. Our approach consists of two main methodological phases. The first phase, is a reduction one that uses the principal component analysis (PCA). As for the second phase, it is based on machine learning and proposes a novel architecture of multilayer perceptrons to provide new values of the inexistant facts. Our experimental study showed promising predictive ability and accepted robustness in the case of a sparse data cube.