Prédiction de valeurs manquantes dans les bases de données — Une première approche fondée sur la notion de proportion analogique

William Correa Beltran, Hélène Jaudoin, Olivier Pivert

Université de Rennes 1 – Irisa, Lannion, France
{William.Correa_Beltran@irisa.fr, Helene.Jaudoin@irisa.fr, Olivier.Pivert@irisa.fr}

Résumé. Cet article présente une méthode originale de prédiction de valeurs manquantes dans les bases de données relationnelles, fondée sur la notion de proportion analogique. Nous montrons en particulier comment un algorithme proposé dans le cadre de la classification automatique peut être adapté à cette fin. Deux cas sont considérés : celui d'une base de données transactionnelle (attributs booléens), et celui où les valeurs manquantes peuvent être de type numérique.

1 Introduction

Dans cet article, nous proposons une solution originale à un problème classique de bases de données qui consiste à prédire/estimer les valeurs manquantes dans une base de données relationnelle incomplète. De nombreuses approches ont été proposées pour traiter cette question, à la fois dans la communauté des bases de données et dans celle de l'apprentissage automatique, fondées sur des dépendances fonctionnelles (Atzeni et Morfuni (1986)), des règles d'association (Ragel (1998)), des règles de classification (Liu et al. (1997)), des techniques de clustering (Fujikawa et Ho (2002)), etc. Nous explorons quant à nous une nouvelle idée, issue de l'intelligence artificielle, qui consiste à exploiter les *proportions analogiques* (Prade et Richard (2012)) pouvant exister dans les données.

La suite de l'article est organisée comme suit. Dans la section 2, nous rappelons les notions de base concernant les proportions analogiques. La section 3 présente le principe général de l'approche que nous proposons pour estimer les valeurs manquantes, inspirée par la technique de classification proposée dans (Bayoudh et al. (2007); Miclet et al. (2008)). La section 4 est consacrée à une expérimentation visant à évaluer les performances de la méthode et à comparer cette dernière avec une technique classique d'estimation (kNN). Finalement, la section 5 rappelle les contributions principales de l'article et trace quelques perspectives de recherche.

2 Rappels sur les proportions analogiques

La présentation qui suit est tirée principalement de Miclet et Prade (2009). Une proportion analogique est une proposition de la forme « A est à B ce que C est à D », ce qui sera noté : $(A : B :: C : D)$. Dans la suite, les objets A, B, C, et D seront supposés être des n-uplets

possédant n attributs, i.e., $A = \langle a_1, \ldots, a_n \rangle, \ldots, D = \langle d_1, \ldots, d_n \rangle$, et nous dirons que A, B, C, et D sont en proportion analogique si et seulement si, pour chaque composante i, une proportion analogique « a_i est à b_i ce que c_i est à d_i » est valide.

Précisons maintenant ce que l'on entend par proportion analogique, en donnant une interprétation intuitive de « est à » et « ce que » dans « A est à B ce que C est à D ». A peut être similaire (ou identique) à B par certains côtés, et différer de B sous d'autres aspects. La manière dont C diffère de D doit être la même que celle dont A diffère de B, tandis que C et D peuvent être similaires à d'autres points de vue, si l'on veut que la proportion analogique soit vérifiée. Cette interprétation est suffisante pour justifier trois postulats remontant à l'époque d'Aristote :

- (ID) $(A : B :: A : B)$
- (S) $(A : B :: C : D) \Leftrightarrow (C : D :: A : B)$
- (CP) $(A : B :: C : D) \Leftrightarrow (A : C :: B : D)$.

(ID) et (S) traduisent la réflexivité et la symétrie de la comparaison « ce que », tandis que (CP) exprime la possibilité de permuter des éléments centraux.

Une *proportion logique* Prade et Richard (2010) est un type particulier d'expression booléenne $T(a,\ b,\ c,\ d)$ faisant intervenir quatre variables a, b, c, d, dont les valeurs de vérité appartiennent à $\mathbb{B} = \{0,\ 1\}$. Elle est composée d'une conjonction de deux équivalences distinctes, mettant en jeu une conjonction des variables $a,\ b$ d'un côté, et une conjonction des variables $c,\ d$ de l'autre côté de $\equiv$, où chaque variable peut être niée. La proportion analogique est un cas particulier de proportion logique et son expression est (cf. Miclet et Prade (2009)) : $(a\overline{b} \equiv c\overline{d}) \wedge (\overline{a}b \equiv \overline{c}d)$. Les six valuations produisant *vrai* sont donc : (0, 0, 0, 0), (1, 1, 1, 1), (0, 0, 1, 1), (1, 1, 0, 0), (0, 1, 0, 1) et (1, 0, 1, 0).

Comme noté dans (Prade et Richard (2013)), l'idée de proportion est fortement liée à celle d'*extrapolation*, autrement dit à l'objectif de deviner/calculer une nouvelle valeur à partir de valeurs existantes, ce qui est bien le but que nous nous fixons ici.

3 Principe de l'approche

3.1 Idée générale

L'approche que nous proposons s'inspire d'une méthode de « classification par analogie » introduite dans (Bayoudh et al. (2007)), où les auteurs décrivent un algorithme appelé FADANA. Ce dernier utilise une mesure de *dissimilarité analogique* entre quatre objets, qui estime dans quelle mesure ces objets sont loin d'être en proportion analogique. En deux mots, la dissimilarité analogique $\mathtt{ad}$ entre quatre valeurs booléennes est le nombre minimal de bits qui doivent être modifiés pour obtenir une analogie valide. Par exemple $\mathtt{ad}(1, 0, 1, 0) = 0$, $\mathtt{ad}(1, 0, 1, 1) = 1$ et $\mathtt{ad}(1, 0, 0, 1) = 2$. Ainsi, en désignant par $\mathcal{A}$ la relation quaternaire de proportion analogique, on a : $\mathcal{A}(a,\ b,\ c,\ d) \Leftrightarrow \mathtt{ad}(a,\ b,\ c,\ d) = 0$.

Lorsque, au lieu d'avoir quatre valeur booléennes, on manipule quatre *vecteurs* booléens dans $\mathbb{B}^n$, il faut ajouter les évaluations $\mathtt{ad}$ obtenues pour chaque composante de façon à obtenir la dissimilarité analogique entre les vecteurs, ce qui conduit à un entier dans l'intervalle $[0, 2n]$. L'algorithme, qui prend en entrée un ensemble d'apprentissage S d'éléments déjà classifiés, un nouvel élément d à classifier, et un entier k, procède comme suit :

1. Pour tout triplet $(a,\ b,\ c)$ de S^3, calcul de $\mathtt{ad}(a,\ b,\ c,\ d)$.

2. Tri de ces n triplets par valeur croissante de leur ad.
3. Si le k^e triplet a la valeur entière p pour ad, alors notons q le plus grand entier tel que le q^e triplet a la valeur p.
4. Résolution des q équations analogiques sur l'étiquette de la classe. On retient le vainqueur des q votes, que l'on affecte comme étant la classe de d.

3.2 Application à la prédiction de valeurs inconnues

3.2.1 Cas des attributs booléens

Cette méthode peut être adaptée au cas de la prédiction de valeurs nulles dans une base de données transactionnelles de la façon suivante. Soit une relation r de schéma $(A_1,\ \ldots,\ A_m)$ et t un n-uplet de r comportant une valeur manquante pour l'attribut A_i : $t[A_i]$ = *NULL*. Pour prédire la valeur de $t[A_i]$ qui est 0 ou 1 dans le cas d'une base de données transactionnelle, on applique l'algorithme précédent en considérant que A_i correspond à la classe cl à déterminer. L'ensemble S d'apprentissage correspond à un échantillon (d'une taille fixée à l'avance) des n-uplets de la relation r (privée de l'attribut A_i qui n'intervient pas dans le calcul de ad mais représente la « classe ») ne comportant aucune valeur manquante. Par ailleurs, on ignore les attributs $A_h,\ h \neq i$ tels que $t[A_h]$ = *NULL* lors du calcul visant à prédire la valeur de $t[A_i]$.

3.2.2 Cas des attributs numériques

Dans le cas d'attributs numériques, une première solution consiste à se ramener au cas booléen : un attribut numérique A est dérivé en autant d'attributs booléens qu'il y a de valeurs dans le domaine actif de A. Cette solution peut cependant paraître discutable car la binarisation conduit à se limiter à des cas d'analogie assez limités (égalité ou non égalité).

Une deuxième solution consiste à rechercher des analogies de type

$$(a : b :: c : d) \Leftrightarrow (\frac{a}{b} = \frac{c}{d}) \tag{1}$$

ou

$$(a : b :: c : d) \Leftrightarrow (a - b = c - d), \tag{2}$$

la première étant appelée proportion *géométrique* et la seconde proportion *arithmétique*.

Ces définitions peuvent être raffinées en introduisant une certaine tolérance sur les relations d'analogie, de façon à couvrir plus de cas. Ainsi on peut considérer que $(100 : 50 :: 80 : 39)$ est *presque vrai* (on a 39 au lieu de 40).

Une façon de procéder est d'adopter une vision graduelle de la dissimilarité analogique, et de considérer que la valeur de AD n'est plus un entier mais un réel. Dans le cas d'une proportion géométrique (formule (1)), pour l'attribut A_i, on peut ainsi ajouter à AD la valeur

$$w_1 = 2 \times \left|1 - \frac{min(ad,\ bc)}{max(ad,\ bc)}\right|.$$

Dans cette formule, la multiplication par 2 a pour but de rendre la pénalité commensurable avec celle appliquée dans le cas booléen (où la dissimilarité analogique peut prendre l'une des valeurs 0, 1 ou 2). Notons que si $max(ad,\ bc) = 0$, cette formule est inapplicable et l'on doit alors se limiter à rechercher une proportion arithmétique (formule (2)).

Dans le cas d'une proportion arithmétique, une solution est de définir, pour chaque attribut numérique, une fonction d'appartenance associée au terme flou « *proche de* ». Pour un attribut donné A_i, on peut définir une fonction trapézoïdale μ_{A_i} de telle sorte que :

$$proche\_de(x,\ y) = \mu_{A_i}(|x-y|) = \begin{cases} 1 \text{ si } |x-y| \leq p_{1,i} \\ 0 \text{ si } |x-y| \geq p_{2,i} \\ \text{linéaire entre les deux.} \end{cases}$$

La pénalité à appliquer se définit alors par : $w_2 = 2 \times (1 - proche\_de(a-b,\ c-d))$.

4 Expérimentation préliminaire

Le principal objectif de l'expérimentation qui a été menée est de comparer les résultats obtenus à l'aide de cette technique avec ceux produits par l'approche classique des plus proches voisins, donc d'estimer son efficacité relative en termes de *précision* (i.e., de pourcentage de valeurs correctement prédites). Nous détaillons ici uniquement des résultats expérimentaux portant sur des bases de données transactionnelles, l'extension au cas d'attributs numériques étant actuellement en cours.

Un ensemble de données de plus de 50 000 n-uplets contenant 20 attributs sur des accidents de la route est utilisé (Geurts et al. (2003)). Un échantillon E est extrait en choisissant 1000 n-uplets de la relation au hasard. Un sous-ensemble M de E est soumis à des modifications, autrement dit, un certain pourcentage de valeurs de chacun de ses n-uplets est remplacé par *NULL*. Ensuite, l'algorithme FADANA est exécuté pour prédire les valeurs manquantes : pour chaque n-uplet d possédant des valeurs manquantes, un échantillon aléatoire D d'une taille fixée de l'ensemble $E - M$ (donc complètes) est choisi. À chaque fois, la méthode des k plus proches voisins (kNN) a été aussi utilisée, en prenant le même nombre de n-uplets et la même valeur de k. Rappelons que la méthode kNN est fondée sur un calcul de distance entre le tuple à compléter et les tuples de l'ensemble d'apprentissage (on ne retient que les k plus proches, et une procédure de vote, analogue à celle de FADANA, permet de prédire la valeur manquante).

On a cherché également à évaluer la proportion de valeurs correctement prédites par FADANA et pas par kNN, et réciproquement. Dans les tableaux qui suivent, No-Fadana (resp. No-kNN) signifie « incorrectement prédit par FADANA (resp. par kNN) », tandis que « Fadana & kNN » signifie « prédit correctement à la fois par FADANA et par kNN ».

TAB. 1 – *Précision en fonction de la valeur de k (proportion de tuples modifiés : 70%, proportion de valeurs modifiées par tuple : 40%, taille de l'ensemble d'apprentissage : 40)*

k	1	5	10	15	30	50	100
FADANA	83,5	84,16	84,33	83,16	84,16	83,83	84,16
kNN	82,83	84,16	83,83	83,66	84	83,33	83,33
FADANA & kNN	74,66	76,66	77,33	76	77	75,83	76,33
FADANA & No-kNN	7	6,16	5,66	6,33	5,5	6,33	5,5
No-FADANA & kNN	6,83	5,83	5	6,66	5,33	5,66	5,66
No-FADANA & No-kNN	9,33	9,33	10,33	10	8,33	9,83	9,66

Le tableau 1 montre comment varie la précision en fonction de la valeur de k utilisée dans l'algorithme. On constate une remarquable stabilité, aussi bien pour FADANA que pour kNN, et l'on peut voir que même avec une valeur de k assez petite, le vote conduit à des résultats corrects dans la grande majorité des cas.

TAB. 2 – *Évolution de la précision en fonction de la taille de l'ensemble d'apprentissage (proportion de tuples modifiés : 70%, $k = 40$, ratio de valeurs modifiées par tuple : 40%)*

$\|App\|$	5	10	20	30	40
FADANA	68,83	82,00	83,66	84,33	84,66
kNN	83,16	84,33	84,50	84,16	83,83
FADANA & kNN	61,16	75,83	76,5	76,5	77,83
FADANA & No-kNN	6,5	4,66	5,83	6,33	5,66
No-FADANA & kNN	20,5	6,83	6,6	6	5
No-FADANA & No-kNN	9,83	10,83	9,33	9,16	10,16

Le tableau 2, quant à lui, montre l'impact que la taille de l'ensemble d'apprentissage a sur la précision. On constate, ce qui n'est guère surprenant, qu'un trop petit ensemble affecte négativement les performances, mais qu'à partir de 30 ou 40 tuples, on atteint pour FADANA un niveau de précision quasiment optimal (autour de 85%). Notons qu'il est illusoire d'espérer atteindre 100% puisqu'il existe en général des tuples qui ne sont en relation de proportion analogique avec aucun triplet de la relation initiale.

Un résultat intéressant est qu'il existe un pourcentage non négligeable (autour de 6%) de valeurs qui sont prédites correctement par FADANA mais pas par kNN, et réciproquement. Ceci permet d'envisager d'utiliser une méthode hybride, qui utiliserait FADANA dans la plupart des cas, mais basculerait vers kNN pour prédire les valeurs dont on peut prévoir qu'elles seront incorrectement estimées par FADANA. L'intuition qu'on peut avoir est que ces dernières se caractérisent par des valeurs élevées de dissemblance analogique `ad` dans la liste construite à l'étape 2 de l'algorithme (voir sous-section 3.1) mais ceci reste à confirmer expérimentalement.

5 Conclusion

Dans cet article, nous avons présenté une méthode originale de prédiction de valeurs manquantes dans les bases de données relationnelles, fondée sur la notion de proportion analogique. Nous avons également montré comment un algorithme proposé dans le cadre de la classification automatique pouvait être adapté à cette fin. Les résultats obtenus, quoique préliminaires, apparaissent encourageants, puisque l'approche conduit à une précision meilleure en moyenne que celle de la technique classique des plus proches voisins.

Il conviendra notamment, dans des travaux futurs, de i) comparer l'approche de prédiction par analogie, au delà de kNN, avec d'autres approches de la littérature ; ii) traiter de façon plus raffinée les attributs catégoriels en prenant en compte des notions telles que synonymie, hyponymie/hypernymie, etc. iii) étudier la façon dont on doit traiter les valeurs prédites lors du processus d'interrogation de la base de données. Cela nécessitera certainement d'utiliser un

modèle de base de données incertaine (par exemple probabiliste), puisque, même si la prédiction a un bon niveau de fiabilité, les valeurs estimées demeurent entachées d'incertitude.

Références

Atzeni, P. et N. M. Morfuni (1986). Functional dependencies and constraints on null values in database relations. *Information and Control 70*(1), 1–31.

Bayoudh, S., L. Miclet, et A. Delhay (2007). Learning by analogy : A classification rule for binary and nominal data. In M. M. Veloso (Ed.), *IJCAI*, pp. 678–683.

Fujikawa, Y. et T. B. Ho (2002). Cluster-based algorithms for dealing with missing values. In M.-S. Cheng, P. S. Yu, et B. Liu (Eds.), *PAKDD*, Volume 2336 of *Lecture Notes in Computer Science*, pp. 549–554. Springer.

Geurts, K., G. Wets, T. Brijs, et K. Vanhoof (2003). Profiling high frequency accident locations using association rules. *Transportation Research Record 1840*, 123–130.

Liu, W. Z., A. P. White, S. G. Thompson, et M. A. Bramer (1997). Techniques for dealing with missing values in classification. In X. Liu, P. R. Cohen, et M. R. Berthold (Eds.), *IDA*, Volume 1280 of *Lecture Notes in Computer Science*, pp. 527–536. Springer.

Miclet, L., S. Bayoudh, et A. Delhay (2008). Analogical dissimilarity : Definition, algorithms and two experiments in machine learning. *J. Artif. Intell. Res. (JAIR) 32*, 793–824.

Miclet, L. et H. Prade (2009). Handling analogical proportions in classical logic and fuzzy logics settings. In C. Sossai et G. Chemello (Eds.), *ECSQARU*, Volume 5590 of *Lecture Notes in Computer Science*, pp. 638–650. Springer.

Prade, H. et G. Richard (2010). Reasoning with logical proportions. In F. Lin, U. Sattler, et M. Truszczynski (Eds.), *KR*. AAAI Press.

Prade, H. et G. Richard (2012). Homogeneous logical proportions : Their uniqueness and their role in similarity-based prediction. In G. Brewka, T. Eiter, et S. A. McIlraith (Eds.), *KR*. AAAI Press.

Prade, H. et G. Richard (2013). Analogical proportions and multiple-valued logics. In *Proc. of the 12th European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty (ECSQARU'13), LNCS, vol. 7958*, pp. 497–509.

Ragel, A. (1998). Preprocessing of missing values using robust association rules. In J. M. Zytkow et M. Quafafou (Eds.), *PKDD*, Volume 1510 of *Lecture Notes in Computer Science*, pp. 414–422. Springer.

Summary

This paper presents an original approach, based on the notion of analogical proportion, to the prediction of null values in a relational database context. We show in particular how an algorithm proposed in the framework of automatic learning can be adapted for this purpose. Two cases are considered: that of a transactional database (where the attributes take Boolean values), and that where the relation handled may involve unknown numerical values.