Classification à partir d'une collection de matrices

Clément Grimal[1], Gilles Bisson[2]

[1] Laboratoire d'Informatique de Grenoble, UMR 5217
Domaine Universitaire de Saint-Martin-d'Hères
38400 Saint Martin d'Hères, France
Clement.Grimal@imag.fr

[2] Laboratoire TIMC, CNRS / UJF 5525
Université de Grenoble - Domaine de la Merci
38710 La Tronche, France
Gilles.Bisson@imag.fr

Résumé : La classification simultanée de deux types d'objets ayant une relation de co-occurence, encore appelée *co-classification*, permet de découvrir la structure (explicite ou latente) de ces objets. Cependant, ce type d'approche ne permet de traiter qu'une seule relation entre deux types d'objets. Or, il existe de nombreux cas où des données multi-relationnelles sont disponibles. Dans le cas des réseaux sociaux par exemple, on possède des données de co-occurrence décrivant les relations entre les acteurs et les différents objets. Dans cet article, nous proposons des extensions de l'algorithme de calcul de co-similarité χ-Sim (Bisson & Hussain, 2008), afin d'exploiter ces données multi-relationnelles.

1 Introduction

L'objectif de la classification est d'organiser un ensemble d'*objets* sous la forme d'un ensemble de classes disjointes ou organisées hiérarchiquement. La similarité d'un couple d'objets appartenant à la même classe doit être maximisée, tandis que celle d'une paire d'objets appartenant à deux classes différentes doit être minimisée. Lorsque les objets que l'on classe sont décrits par des variables qui sont elles-mêmes classifiables, par exemple des documents décrits par des mots, on peut utiliser des méthodes de classification simultanée ou « co-classification ». Cependant, ces méthodes ne s'applique qu'à deux types d'objets, et ne permettent donc pas de tirer partie de la richesse de données multi-relationnelles. Nous proposons donc d'exploiter ces informations supplémentaires pour améliorer la classification des objets qu'elles décrivent. Des méthodes

Ce travail a bénéficié d'une aide de l'Agence Nationale de la Recherche portant la référence ANR-08-CORD-009, dans le cadre du projet FRAGRANCES.

similaires ont été proposées par (Long *et al.*, 2006; Banerjee *et al.*, 2007). L'objectif de notre méthode est de calculer les mesures de similarités entre les objets du même type, en utilisant toutes les données disponibles, le travail se base sur l'algorithme de calcul de co-similarité χ-Sim (Bisson & Hussain, 2008). Nous traiterons en tant qu'exemple le problème de classification de films selon leur genre, chaque film étant décrit par un ensemble d'acteurs, et un ensemble de mots-clés.

2 Algorithmes

2.1 Notations utilisées

Matrices de données : soit $\mathbf{M_p}$ l'ensemble des matrices contenant les informations de la base de données. Chaque matrice comporte r_p lignes et c_p colonnes : chaque valeur $m_{p,ij}$ caractérise *la nature de la relation* entre le $j^{ème}$ acteur (ou mot-clé) et le $i^{ème}$ film ; $\mathbf{m_{p,i}} = \left[m_{p,i1} \cdots m_{p,ic_p}\right]$ est le vecteur ligne correspondant au film i et $\mathbf{m_p^j} = \left[m_{p,1j} \cdots m_{p,r_pj}\right]$ est le vecteur colonne correspondant à l'acteur (ou mot-clé) j.

Matrices de similarité : $\mathbf{SR_p}$ (de taille $r_p \times r_p$) et $\mathbf{SC_p}$ (de taille $c_p \times c_p$) représentent les matrices carrées et symétriques contenant respectivement les valeurs de similarité entre films et entre acteurs (ou entre mots-clés) avec $sr_{p,ij} \in [0,1]$, $1 \leq i,j \leq r_p$ et $sc_{p,ij} \in [0,1]$, $1 \leq i,j \leq c_p$.

2.2 L'algorithme de co-similarité χ-Sim

L'algorithme χ-Sim (Bisson & Hussain, 2008) ne travaillant que sur une seule matrice $\mathbf{M_p}$ à la fois, nous pouvons simplifier les notations précédentes pour supprimer les indices p. Nous considérons que $\mathbf{M}$ est une matrice de co-occurrence *films-acteurs*, et nous cherchons à calculer la matrice de similarité entre *films* $\mathbf{SR}$.

Le principe de l'algorithme est de considérer simultanément que deux films se ressemblent s'ils contiennet des acteurs similaires en se basant sur la matrice $\mathbf{SC}$ (et non strictement les acteurs identiques comme c'est cas pour les mesures classiques). On effectue de manière duale le même calcul pour comparer chaque paire d'acteurs.

Le problème s'exprime alors sous la forme d'un système d'équations permettant de calculer $\mathbf{SR}$ et $\mathbf{SC}$, et pouvant s'écrire comme produit de trois matrices. Ainsi, $\mathbf{SR}$ et $\mathbf{SC}$) se calcule de cette manière [1] :

$$\mathbf{SR} = \mathbf{M} \times \mathbf{SC} \times \mathbf{M}^T \circ \mathbf{NR} \quad \text{avec } nr_{ij} = \frac{1}{|\mathbf{m_i}| \times |\mathbf{m_j}|} \tag{1}$$

$$\mathbf{SC} = \mathbf{M}^T \times \mathbf{SR} \times \mathbf{M} \circ \mathbf{NC} \quad \text{avec } nc_{ij} = \frac{1}{|\mathbf{m^i}| \times |\mathbf{m^j}|} \tag{2}$$

Les équations décrites dans (1 et 2) n'étant pas indépendantes, on utilise une méthode itérative pour calculer $\mathbf{SR}$ et $\mathbf{SC}$. Les deux matrices de similarité sont initialisées avec la matrice identité $\mathbf{I}$, représentant l'absence de connaissance a priori. Puis, on calcule

1. L'opérateur « $\circ$ » désigne le produit de Hadamard.

tour-à-tour $\mathbf{SR}$ en utilisant la matrice $\mathbf{SC}$ de l'itération précédente, et $\mathbf{SC}$ en utilisant la matrice $\mathbf{SR}$ de l'itération précédente. Il est intéressant de remarquer que l'itération n permet d'explorer un chemin de longueur n dans le graphe bi-partite associé au jeu de données.

2.3 Extension aux données multi-relationnelles

On cherche maintenant à utiliser l'algorithme précédent permettant de calculer les co-similarité de deux types d'objets, pour calculer des co-similarité de différents types d'objets.Nous présentons ici une méthode utilisant une structure en anneau. Elle consiste à réaliser une seule itération avec chacune des matrices de données, chaque matrice de similarité obtenue à l'aide de χ-Sim alimentant la seconde instance de l'algorithme. On alterne ainsi entre deux « vues » sur les données, ce qui permet de faire circuler les informations de similarité de l'une à l'autre.

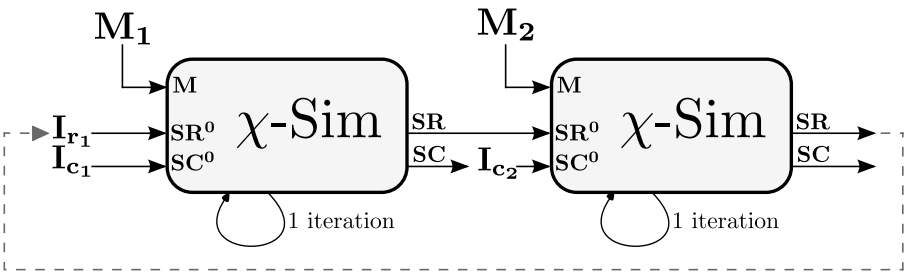

FIGURE 1 – Schéma de la méthode de structure en anneau. $\mathbf{I_m}$ représente la matrice identité de dimension $\mathbf{m}$. La ligne pointillée signifie que l'on peut effectuer une nouvelle itération utilisant $\mathbf{M_1}$, etc.

3 Expérimentations et résultats

3.1 Dispositif expérimental

En utilisant *IMDb* [2], nous avons créer notre jeu de données comportant 617 films de 17 genres différents (environ 36 films par genre), 1878 mots-clés et 1398 acteurs, représenté par deux matrices : films - mots-clés ($\mathbf{M_1}$) et films - acteurs ($\mathbf{M_2}$). L'objectif est ensuite de classer ensemble les films du même genre. Nous avons comparé nos algorithmes à la mesure de similarité de cosinus, à la mesure de similarité induite par LSA (Latent Semantic Analysis) de (Deerwester *et al.*, 1990) et à la mesure de co-similarité χ-Sim. Nous avons retenu l'algorithme de Classification Ascendante Hiérarchique (CAH) avec l'indice de Ward pour construire les classes de films. Pour évaluer les résultats obtenus, nous avons utilisé l'indice classique de « précision micro-moyennée » ou micro-averaged Precision (Pr) introduit par (Dhillon *et al.*, 2003).

2. http://www.imdb.com/interfaces/

3.2 Résultats expérimentaux

Pour LSA, le nombre de valeurs singulières conservées est celui maximisant la précision micro-moyennée. Et pour χ-Sim, seuls les meilleurs résultats sont affichés.

	Cosinus	LSA	χ-Sim		Anneau
			3 it.	4 it.	$\mathbf{M_1} \leftrightarrow \mathbf{M_2}$
$\mathbf{M_1}$	0,225	0,277	0,282	0,285	3 it. : **0,292**
$\mathbf{M_2}$	0,212	0,216	0,217	0,220	4 it. : 0,219

TABLE 1 – Résultats des expérimentations en terme de précision micro-moyennée (Pr) utilisant les méthodes classiques, et la structure en anneau.

Le tableau 1 présente les résultats obtenus avec les méthodes classiques de mesures de (co-)similarité, et la méthode utilisant la structure en anneau. Le meilleur résultat obtenu est donc celui utilisant la structure en anneau, avec **0,292** de précision micro-moyennée.

4 Conclusion

Dans cet article, nous avons présenté les premiers résultats d'une méthode permettant de tirer partie des informations fournies par des jeux de données multi-relationnels. Notre méthode s'appuie sur l'algorithme de calcul de co-similarité χ-Sim développée par (Bisson & Hussain, 2008). Les différentes « instances » de l'algorithme χ-Sim que notre méthode utilise, correspondent à plusieurs vues sur les données dont nous disposons, agréger correctement ses vues permet ainsi d'améliorer la qualité des résultats de classification. Nos futurs travaux incluent l'expérimentation d'autres extensions, ainsi que la généralisation de χ-Sim à plus de deux dimensions.

Références

BANERJEE A., BASU S. & MERUGU S. (2007). Multi-way clustering on relation graphs. In *SDM* : SIAM.

BISSON G. & HUSSAIN F. (2008). Chi-sim : A new similarity measure for the co-clustering task. In *Machine Learning and Applications, 2008. ICMLA '08. Seventh International Conference on*, p. 211–217.

DEERWESTER S., DUMAIS S. T., FURNAS G. W., THOMAS & HARSHMAN R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, **41**, 391–407.

DHILLON I. S., MALLELA S. & MODHA D. S. (2003). Information-theoretic co-clustering. In *Proceedings of The Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining(KDD-2003)*, p. 89–98.

LONG B., WU X., ZHANG Z. M. & YU P. S. (2006). Unsupervised learning on k-partite graphs. In *KDD '06 : Proceedings of the 12th ACM SIGKDD*, p. 317–326, New York, NY, USA : ACM.