Transformation de l'espace de description pour l'apprentissage par transfert

Nistor Grozavu *, Younès Bennani*, Lazhar Labiod**

*LIPN-UMR 7030, Université Paris 13,
99, av. J-B Clément, 93430 Villetaneuse, France
email: {firstname.secondname}@lipn.univ-paris13.fr
**LIPADE, Paris Descartes University
45, rue des Saints Pères, 75006 Paris, France
email: {lazhar.labiod}@parisdescartes.fr

Résumé. Dans ce papier, nous proposons une étude sur l'utilisation de l'apprentissage topologique pondéré et les méthodes de factorisation matricielle pour transformer l'espace de représentation d'un jeu de données "sparse" afin d'augmenter la qualité de l'apprentissage, et de l'adapter au cas de l'apprentissage par transfert. La factorisation matricielle nous permet de trouver des variables latentes et l'apprentissage topologique pondéré est utilisé pour détecter les plus pertinentes parmi celles-ci. La représentation de nouvelles données est basée sur leurs projections sur le modèle topologique pondéré.

Pour l'apprentissage par transfert, nous proposons une nouvelle méthode où la représentation des données est faite de la même manière que dans la première phase, mais en utilisant un modèle topologique élagué.

Les expérimentations sont présentées dans le cadre d'un Challenge International où nous avons obtenu des résultats prometteurs (5ieme rang de la compétition internationale).

1 Introduction

L'exploration des données, un domaine en pleine évolution et interdisciplinaire, a reçu beaucoup d'intérêt dans de nombreux domaines scientifiques. L'objectif de l'exploration de données est d'extraire des connaissances à partir d'ensembles de données volumineux en combinant les méthodes de statistique et d'intelligence artificielle avec les méthodes de la gestion de bases de données.
La taille des données peut être mesurée selon deux dimensions, le nombre de variables et le nombre d'observations. Ces deux dimensions peuvent prendre des valeurs très élevées, ce qui peut poser un problème lors de l'exploration et l'analyse de ces données. Pour cela, il est fondamental de mettre en place des outils de traitement de données permettant une meilleure compréhension des données.
Dans cette étude, nous nous intéressons à la réduction de dimension de l'espace de description dans le cadre de l'apprentissage non-supervisé à travers la factorisation matricielle et la

transformation de cet espace afin de faciliter le processus d'apprentissage par transfert. La factorisation approximative et la factorisation tensorielle (ou la décomposition) d'une matrice jouent un rôle fondamental dans l'amélioration des données et l'extraction de composantes latentes. Un point commun pour la suppression du bruit, la réduction du modèle, la reconstruction de faisabilité, et la séparation aveugle de sources (Blind Source Separation) est de remplacer les données d'origines par une représentation approximative reduite des dimensions obtenues via une matrice, ou une factorisation ou éventuellement d'une décomposition matricielle. La décomposition en valeurs singulières (SVD) traite les lignes et les colonnes d'une manière symétrique, et fournit donc plus d'informations sur la matrice des données. Cette méthode permet aussi de trier l'information contenue dans la matrice de telle sorte que, de façon générale, la «partie pértinente» devienne visible. C'est la propriété qui rend la SVD si utile en "data mining" et de nombreux autres domaines. La méthode de bidiagonalisation GK (Golub-Kahan) a été initialement formulée (Golub et Kahan, 1965) pour le calcul de la SVD. Cette méthode peut être aussi utilisée pour calculer une bidiagonalisation partielle.

Dans notre méthode nous utilisons cette technique pour les données "sparse" et l'Analyse en Composants Principales (ACP) pour les autres jeux de données.

Le reste de cet article est organisé comme suit. La Section 2 présente brièvement le principe de la factorisation matricielle et l'utilisation de cette technique pour le clustering, ainsi que les principes de l'apprentissage par transfert. Les methodes proposées pour l'apprentissage non supervisé et l'apprentisage topologique par transfert sont presentées dans la section 2.1 et 2.2 respectivement. Dans la section 3, nous présentons les résultats de la validation et leur interprétation. Une conclusion et des perspectives sont données dans la section 4.

2 Transformation de l'espace de description

2.1 Transformation non supervisée

L'apprentissage non-supervisé est souvent utilisé pour le clustering des données et rarement comme un procedé de prétraitement de données. Toutefois, il existe un certain nombre de méthodes qui produisent de nouvelles représentations de données à partir des données non étiquetés. Ces méthodes non supervisées sont parfois utilisées comme un outil de prétraitement pour des modèles d'apprentissage supervisé. Étant donné une matrice de données représentée comme des vecteurs de variables (p observations en lignes et n caractéristiques en colonnes), le but de la transformation non supervisée de l'espace de description est de produire une autre matrice de données de dimension (p, n') (la représentation transformée de n' nouvelles variables latentes) ou une matrice de similarité entre les données de dimension (p, p). L'application d'une méthode supervisée sur la matrice transformée doit fournir de meilleurs résultats par rapport à la base de données initiale. La transformation de l'espace de description se fait suivant deux étapes successives. Dans un premier temps, nous décomposons la matrice "sparse" des données selon la méthode SVD. Ensuite, la matrice des variables latentes obtenue après cette décomposition est utilisée pour l'apprentissage d'un modèle topologique de type lwo-SOM (Grozavu et al. (2009)), qui permet de détecter et de pondérer les caractéristiques pertinentes. Le codage finale de chaque donnée est basé sur les distances de chaque donnée à chaque prototype du modèle lwo-SOM. Cette dernière matrice des distances représente la nouvelle description des données. Pour évaluer la qualité de ce nouveau codage des données,

la nouvelle représentation est présentée par la suite à un classificateur de type Analyse Discriminante Linéaire (ADL).

Pour une base de données d'apprentissage A, une base d'évaluation B, et une base Finale (de test), C, voici la méthode proposée pour la transformation de l'espace de description :

1. Normalisation : $\widehat{A} = A * diag(std(A))^{\frac{1}{2}}$
2. Réduction de la dimension de $\widehat{A}$ par la factorisation matricielle : $[U_{\widehat{A}} S_{\widehat{A}} V_{\widehat{A}}] = svd(\widehat{A})$
   Pour chaque colonne de $U_{\widehat{A}}$, $U_k = \frac{U_k}{\|U_k\|}$
3. Quantification matricielle : $P_k = lwo - SOM(U_{\widehat{A}})$ où k est le nombre de vecteurs propres retenus
4. Appliquer l'étape 1 et 2 sur les matrices B et C, $[U_{\widehat{B}} S_{\widehat{B}} V_{\widehat{B}}] = svd(\widehat{B})$
   $[U_{\widehat{C}} S_{\widehat{C}} V_{\widehat{C}}] = svd(\widehat{C})$
5. Calcul des matrices de distances : $D = (d_{ij})$ où $d_{ij} = \|U_i - P_j\|^2$

2.2 Transformation semi-supervisée

Les modèles prédictifs capables de classer de nouvelles instances (prédire correctement les étiquettes) nécessitent généralement un apprentissage en utilisant de grandes quantités de données labellisées.
Malheureusement, peu de données étiquetées d'apprentissage peuvent être disponibles en raison du coût de l'annotation manuelle des données. Dans certains cas pratiques, il est souhaitable de produire des représentations de données qui peuvent être réutilisables d'un domaine à un autre. Dans cette étude, nous voulons examiner comment une représentation développée avec un ensemble d'étiquettes peut être utilisée pour apprendre d'une manière plus facile une nouvelle tâche similaire ou proche. Par exemple, dans le domaine de la reconnaissance d'écriture, les chiffres manuscrits étiquités seront disponibles pour l'apprentissage. La tâche d'évaluation serait alors la reconnaissance de lettres manuscrites alphabétiques. Nous appelons ce type d'apprentissage «apprentissage par transfert".
Pour l'apprentissage par transfert, nous proposons une nouvelle méthode de transformation de l'espace de description où la représentation des données est faite de la même manière que dans la première méthode non supervisée, mais en utilisant une carte $lwo - SOM$ élaguée. Cet élagage est effectué après l'étiquetage de la matrice des prototypes lwo-SOM en utilisant les étiquettes disponibles. L'élagage consiste à éliminer tous les prototypes étiquetés (qui representent les données étiquetés) et on obtient donc une décomposition de la matrice initiale, ayant comme resultat la matrice de prototypes non étiquetés. En effet, cette nouvelle matrice, represente les données des autres classes qui ne sont pas disponibles pour le transfert. Ces prototypes seront utilisés comme un dictionnaire pour le codage des données de validation et d'évaluation finale.

En effet, les ensembles des données de validation et d'évaluation finale sont projetés sur les prototypes non étiquetés en calculant la distance Euclidienne entre ces observations et les prototypes de la carte lwo-SOM. Cette dernière matrice des distances représentera le nouvel espace de description (figure 1).

Transformation de l'espace de description

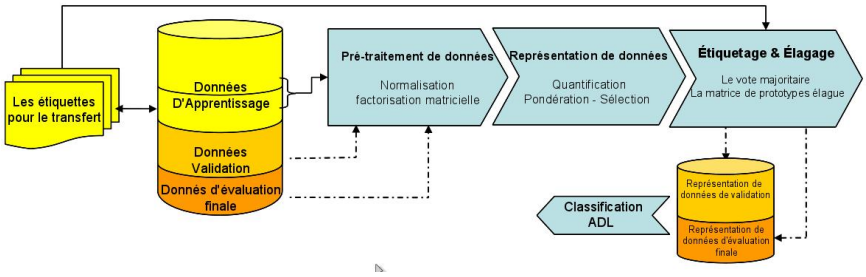

FIG. 1 – *Transformation de l'epace de description et l'Apprentissage par transfert*

3 Validation des approches proposées

3.1 Protocole expérimental

Les deux méthodes pour la transformation de l'espace des données que nous proposons dans ce travail ont été testées dans le cadre d'un Challenge International sur L'Apprentissage non supervisé et par Transfert (Unsupervised and Transfer Learning Challenge) Guyon et al. (2011). Le Challenge a été constitué de deux phases : L'apprentissage non supervisé pour la transformation de l'espace de données et l'apprentissage par transfert. Plus de détails concernant le Challenge peuvent être trouvés sur le site web officiel du Challenge
(http ://www.causality.inf.ethz.ch/unsupervised-learning.php).
Dans la première phase du challenge, aucune étiquette n'est fournie aux participants. Les participants sont invités à produire des représentations de données qui seront évaluées par les organisateurs sur les tâches d'apprentissage supervisé.

Les étiquettes des tâches d'apprentissage supervisé utilisé à des fins d'évaluation resteront inconnues pour les participants à la phase 1 et 2, mais d'autres labels seront disponibles pour l'apprentissage par transfert dans la phase 2.
Dans la deuxième phase du challenge (l'apprentissage par transfert), certaines étiquettes sont fournies aux participants pour les mêmes ensembles de données utilisés dans la première phase, ce qui normalement va permetre d'améliorer les représentations de données obtenues dans la prémière phase.
Cinq bases de données ont été mises à la disposition des participants au Challenge. Le tableau 1 résume une description des ensembles des données utilisées pour valider nos approches.

Les performances de prédiction sont évaluées en fonction de la AUC et de l'aire sous la courbe d'apprentissage (ALC) sur l'ensemble de test versus le nombre d'exemples utilisés pour reéaliser l'apprentissage. L'AUC (Area Under the ROC Curve) Fawcett (2004) est calculeée pour toutes les observations du jeu de donneées Salperwyck et Lemaire (2011).

Pour le jeu de données AVICENNA, nous obtenons un petit score AUC de 0.15 et 0.18 , mais cela est normal vu que c'est un problème assez difficile pour l'apprentissage non supervisé. Pour la base de données HARRY, nous avons construit une matrice de prototypes de taille 900 (30x30 cellules) avant de transformer le jeu de données initial en utilisant une factorisation matricielle de type SVD (20 vecteurs propres). Nous avons utilisé la SVD pour la base RITA,

TAB. 1 – *Les bases de données*

Dataset	Domain	Var.	Spars.	App.	Transf.
AVICENNA	Handwriting	120	0%	150205	50000
HARRY	Video	5000	98.1%	69652	20000
RITA	Images	7200	1.1%	111808	24000
SYLVESTER	Ecology	100	0%	572820	100000
TERRY	Text	47236	99.8%	217034	40000

et nous avons construit une matrice de prototypes de taille 900 (30x30 cellules). Les résultats obtenus nous ont permis de se positionner en deuxième rang pour ce jeu de données dans le Challenge. Après la réduction de la dimension du jeu de données SYLVESTER en utilisant l'ACP, nous avons construit la matrice des prototypes de taille 1600 (40x40 cellules) et nous avons obtenu le score AUC de 0.45 pour la base d'évaluation finale. Finallement, pour la base de données TERRY avec 47236 variables, nous avons utilisé la méthode lwo-SOM pour obtenir une matrice de prototypes de taille 1089 (carte 33x33) après une transformation matricielle de la base initiale en utilisant la SVD.

TAB. 2 – *Les résultats experimentaux pour l'Apprentissage non supervisé*

Base de données	App. non supervisé		App. par transfert	
	AUC	ALC	AUC	ALC
avicenna	0.701728	0.182106	0.623894	0.105119
harry	0.961722	0.709893	0.961722	0.709893
rita	0.786303	0.489439	0.759892	0.363303
sylvester	0.825077	0.44926	0.624744	0.126217
terry	0.994574	0.808953	0.888154	0.566029

Les résultats du Challenge peuvent être consultés sur le site oficiel du Challenge (notre équipe a le nom NG-A3) :

`http://www.causality.inf.ethz.ch/unsupervised-learning.php?page=results#cont`

En analysant les résultats du Challenge, nous pouvons conclure que l'approche que nous avons proposée offre des performances qui dépassent largement d'autres méthodes comme : celles basées sur les Forêts Aléatoires ; la factorisation non-negative ; l'analyse factorielle ; la réduction de la dimensionalité avec RBM, ...

Par contre, dans la première phase du Challenge, le gagnant (nom de l'équipe : AIO) Guyon et al. (2011) ont utilisé un algorithme d'apprentissage à base de noyaux. En utilisant les données de validation, ils ont progressivement amélioré le noyau.

4 Conclusion

Dans ce travail, nous avons proposé deux méthodes de transformation de l'espace de description des données. Une méthode basée sur la combinaison d'une technique de décomposition matricielle et d'une classification topologique pondérée, et une extension qui utilise un processus semi supervisé pour l'élagage du modèle topologique. Nous avons adapté ces méthodes pour le Challenge "Unsupervised and Transfer Learning" afin de transformer l'espace des caractéristiques des différents jeux de données. Nos approches ont démontré une grande efficacité pour des problèmes de grandes dimensions et de différent types de données. Pour la deuxième phase du Challenge, une méthodologie d'Apprentissage par Transfert de nouvelles connaissances a été proposée en utilisant une technique d'élagage de la matrice des prototypes obtenue avec lwo-SOM. Les résultats obtenus sont très prometteurs et ouvrent de nouvelles perspectives.

Références

Fawcett, T. (2004). Roc graphs : Notes and practical considerations for researchers. *Machine Learning*.

Golub, G. H. et W. Kahan (1965). Calculating the singular values and pseudo-inverse of a matrix. *in SIAM J. Numer. Anal.*, 205–224.

Grozavu, N., Y. Bennani, et M. Lebbah (2009). From variable weighting to cluster characterization in topographic unsupervised learning. In *in Proc. Proc. of IJCNN09, International Joint Conference on Neural Network*.

Guyon, I., G. Dror, V. Lemaire, G. Taylor, et D. W. Aha (2011). Unsupervised and transfer learning challenge. In *Proc. of International Joint Conference on Neural Networks 2011*.

Salperwyck, C. et V. Lemaire (2011). Impact de la taille de l'ensemble d'apprentissage : une étude empirique. *Conférence Internationale Francophone sur l'Extraction et la Gestion de Connaissance*.

Summary

In this paper, we propose a study on the use of toplogical weighting learning to transform the representation space of a dataset in order to increase the quality of learning and adapting it to the case of Transfer Learning. For the transformation of the feature space, we use weighted topological models (lwo-SOM) and PCA or SVD to find a new space of representation. New data representation is based on their projections on the topological weighted model. Each example in the dataset is described by a new representation consisting of the distances of this example to all components of the topological model (prototype). For the Transfer Learning , we propose a new method where the representation of data is done in the same way as in the first phase, but using a pruned topological model. The experiments are presented as part of an International Challenge where we obtained good results (5th place).