Régression Laplacienne semi-supervisée pour la reconstitution des dates de pose des réseaux d'assainissement

Vivien Kraus *, Khalid Benabdeslem *,
Frederic Cherqui **

*Université Lyon 1 - 43 Bd du 11 Novembre 1918, 69622 Villeurbanne
**Univ Lyon, INSA-LYON, Université Claude Bernard Lyon 1,
DEEP, F-69621, F-69622, Villeurbanne

Résumé. La date de pose est souvent un facteur principal d'explication de la dégradation des conduites d'assainissement. Pour les gestionnaires de ces réseaux, connaître cette information permet ainsi (par l'utilisation de modèles de détérioration) de prédire l'état de santé actuel des conduites non encore inspectées. Cette connaissance est primordiale pour prendre des décisions dans un contexte de forte contrainte budgétaire. L'objectif est ainsi de reconstituer ces dates de pose à partir des caractéristiques du patrimoine et de son environnement. Les données à manipuler présentent plusieurs niveaux de complexité importants. Leurs sources sont hétérogènes, leur volume est important et les informations sur leur étiquetage (dates) sont limitées : seulement 24 % du linéaire est connu pour les réseaux d'assainissement de la métropole de Lyon. La base de données sous-jacente contient les caractéristiques connues des conduites (profil géométrique, matériau utilisé, etc.). Dans ce papier, nous proposons de mesurer l'effet et l'impact de quelques méthodes d'apprentissage statistique semi-supervisé, et de proposer ainsi une approche alternative adaptée à la reconstitution de ce type de données.

1 Introduction

Depuis la prolifération des bases de données partiellement étiquetées, l'apprentissage automatique a connu un développement important dans le mode semi-supervisé [Chapelle et al. (2006)]. Cette tendance est due à la difficulté de l'étiquetage des données d'une part et au coût de cet étiquetage quand il est possible, d'autre part. L'apprentissage semi-supervisé est un cas particulier de l'apprentissage à partir de données faiblement étiquetées [Li et al. (2013)], qui consiste en général à modéliser une fonction statistique à partir de données regroupant à la fois des exemples étiquetés et d'autres non-étiquetés. Pour aborder une telle problématique, deux grandes familles d'approches existent : celle basée sur la propagation de la supervision en vue de l'apprentissage supervisé [Zhu (2006)] et celle basée sur la transformation de la partie étiquetée en contraintes en vue de leur intégration dans un processus de clustering (non-supervisé) [Basu et al. (2008)]. Nous nous intéressons ici à la première famille d'approches avec une difficulté particulière. Il s'agit d'apprendre avec une partie supervisée relativement

réduite par rapport à la partie non-supervisée. Dans ce paradigme semi-supervisé, la littérature a connu un essor important, depuis déjà une vingtaine d'années, notamment en classification, avec des approches populaires comme le self-training [Chapelle et al. (2006)], le co-training [Blum et Mitchell (1998)], Transductive-SVM ou S^3VM [Joachims (1999), Bennett et Demiriz (1999)], les approches à base de graphes [Blum et Chawla (2001)] et les approches génératives [Nigam et al. (2000)]. Dans ce même paradigme, les problèmes de régression ont également suscité l'intérêt de plusieurs travaux de recherche, que nous pouvons citer sans vouloir être exhaustifs. Il s'agit d'approches diverses : à base de régression linéaire [Azriel et al. (2016); Ji et al. (2012)], logistique [Amini et Gallinari (2002)] ou Laplacienne [Cai et al. (2006), Belkin et al. (2006)] ; d'autres utilisant le principe du co-training [Zhou et Li (2005)] ; ou satisfaisant des contraintes plus spécifiques liées à l'ordre de préférences entre les données non-étiquetées [Zhu (2006)] ou à leur distribution géométrique dans les espaces multidimensionnels [Ryan et Culp (2015), Moscovich et al. (2016)].

Dans ce papier, nous proposons d'améliorer l'algorithme proposé par [Ji et al. (2012)]. L'idée est de bénéficier de la réduction de dimensions proposée dans leur travail et de remplacer la régression linéaire par une régression Laplacienne qui est censée reconnaître mieux la structure géométrique induite des données non-étiquetées. Pour ce faire, nous nous appuyons en plus sur les travaux de [Belkin et al. (2006)]. Ces deux approches seront donc à la base de l'approche proposée et seront décrites dans la section suivante. Ensuite, nous appliquons cette proposition sur des données des réseaux d'assainissement de la métropole de Lyon pour la reconstitution des dates de pose des canalisations. En effet, cette connaissance est primordiale pour les collectivités locales afin d'inspecter l'état de santé de ces réseaux. L'approche retenue ne prend pas en compte la dimension spatiale de l'évolution (temporelle) des réseaux car cette approche fait l'objet de travaux spécifiques en lien avec un géographe [1].

2 Notations et formulations

Notons par : n le nombre total d'individus, m le nombre d'individus labellisés, $(\mathbf{v_i})_{i=1}^{n}$ les données d'entrée, $\mathbf{z}$ la variable cible, de dimension n (dont seulement m éléments ne sont pas manquants). Dans notre cas, les données d'entrée sont dans un espace multidimensionnel $\mathcal{X}$ et la cible est réelle. Notons également $\kappa\colon \mathcal{X} \to \mathcal{X}$ le noyau de Mercer [Mika et al. (1999)] que l'on utilise pour tenir compte des non-linéarités de la régression. On note K la matrice carrée symétrique réelle, définie pour $(i,j) \in \{1,...,n\}^2$ par : $K_{i,j} = \kappa(\mathbf{v_i}, \mathbf{v_j})$. Cette matrice est potentiellement très grande par son nombre d'éléments (n^2). Dans notre application nous prendrons le noyau RBF : $\kappa(\mathbf{v_i}, \mathbf{v_j}) = e^{-\frac{\|\mathbf{v_i}-\mathbf{v_j}\|^2}{2\omega^2}}$, qui rajoute un hyperparamètre ω. L'extraction des s composantes principales de K de plus hautes valeurs propres définit les données d'apprentissage dans une nouvelle matrice X de n lignes par s colonnes. La régularisation Laplacienne fait intervenir la matrice Laplacienne d'un certain graphe L, dont la matrice d'adjacence en i,j est 1 si i et j sont ressemblantes (la distance dans l'espace des variables est inférieure à un seuil `radius`), 0 sinon, ainsi que deux régulariseurs τ et C. Enfin, les données de test sont notées $(\mathbf{v_{ti}})_{i=1}^{b}$.

1. `http://www.hireau.org`

3 Approche proposée : LapS3L

Nous proposons une méthode, que nous appelons LapS3L, qui consiste à améliorer l'algorithme SSSL [Ji et al. (2012)], tout en remplaçant la deuxième étape de l'algorithme par la régression semi-supervisée LapRLS [Belkin et al. (2006)] dans le cas linéaire. Il y a de nombreuses motivations pour cela :

1. La première étape du SSSL est censée traiter le problème de la *non linéarité*, on peut donc espérer n'avoir à faire qu'une régression linéaire dans la seconde étape ;
2. La régularisation Laplacienne est une généralisation des moindres carrés : si les paramètres ν et γ_{lap} sont nuls, alors le problème linéaire résolu avec la régularisation Laplacienne est exactement le même que le problème des moindres carrés utilisés pour la deuxième étape du SSSL.

L'approche que nous proposons est légèrement différente du principe déjà évoqué dans [Chapelle et al. (2006)] et mis en place dans le SSSL (décomposer le problème semi-supervisé en une étape non supervisée puis une étape complètement supervisée). La deuxième étape est aussi semi-supervisée, mais elle utilise des données que l'on peut traiter par une méthode linéaire. Par rapport au SSSL, on introduit une régularisation dans le calcul de w (ligne 9 dans l'algorithme 1).

Algorithm 1 LapS3L

1: **procedure** LEARNING($(\mathbf{v_i})_{i=1}^{n}, (z_i)_{i=1}^{n}, \kappa, s, \boldsymbol{\tau}, \boldsymbol{C}$)
2: $\quad\forall i,j \in \{1,...,n\}, K_{i,j} \leftarrow \kappa(\mathbf{v_i}, \mathbf{v_j})$
3: $\quad$Trouver $(\sigma_i, \mathbf{u_i})_{i=1}^{s}$, les s plus fortes valeurs propres et vecteurs propres associés de K
4: $\quad D \leftarrow \mathrm{diag}(\sigma_i)_{i=1}^{n}$
5: $\quad\forall i \in \{1,...,n\}, \forall j \in \{1,...,s\}, U_{i,j} \leftarrow \mathbf{u_j}_i$ ▷ Chaque vecteur est une colonne
6: $\quad X \leftarrow {}^tKU$
7: $\quad$Trouver $\left(i_k^{\mathrm{labeled}}\right)_{k=1}^{m}$, les indices des individus labellisés
8: $\quad\forall k \in \{1,...,m\}, X_{lk,.} \leftarrow X_{i_k^{\mathrm{labeled}},.}, z_{lk} \leftarrow z_{i_k^{\mathrm{labeled}}}$
9: $\quad\boldsymbol{w \leftarrow \left[\tau^t XLX + {}^tX_lX_l + CI_s\right]^{-1} {}^tX_lz_l}$
10: $\quad\gamma \leftarrow D^{\frac{1}{2}}w$
11: $\quad$**procedure** PREDICTION($(\mathbf{v_{ti}})_{i=1}^{b}$)
12: $\quad\quad\forall i \in \{1,...,n\}, \forall j \in \{1,...,b\}, K_{bi,j} \leftarrow \kappa(\mathbf{v_i}, \mathbf{v_t j})$
13: $\quad\quad\Phi \leftarrow D^{-\frac{1}{2}}{}^tUK_b$
14: $\quad\quad\hat{z} \leftarrow \Phi^t\gamma^*$
15: $\quad$**end procedure**
16: **end procedure**

4 Application aux données d'assainissement

Les réseaux d'assainissement ont été construits et étendus pour et par la ville. Ce patrimoine existant impacte les pratiques de gestion : de nombreuses études [Ahmadi et al. (2014), Harvey et McBean (2014)] ont montré l'importance primordiale de la connaissance de la date

de pose des conduites pour estimer leur état actuel de détérioration et prédire leur dégradation. L'enjeu est donc pour la Métropole de Lyon de reconstituer les dates de pose des réseaux d'assainissement dont seulement 24 % du linéaire est connu.

4.1 Description des données

La base de données contient sept variables, dont quatre variables catégorielles (matériau, forme, structurant, type d'effluent) et trois variables continues (longueur, largeur, hauteur). Pour représenter les données catégorielles, nous utilisons un encodage où chaque modalité de chaque variable est représentée par une variable booléenne. La variable `Forme` possède plus de 300 modalités, ce qui porte la dimension des données à 346. Le choix des variables s'est fait en concertation avec les experts métiers de la Direction de l'Eau de la Métropole de Lyon. Les individus sont les conduites du réseau d'assainissement. Le réseau contient 85766 conduites, mais nous n'en gardons que 4000 (pour pouvoir traiter la matrice K). Enfin, la variable cible est la date de pose. Elle est comprise entre 1900 et 2014 (avant normalisation).

4.2 Résultats

En appliquant notre méthode d'apprentissage semi-supervisé à la base de données Assainissement, on obtient une erreur quadratique moyenne RMSE de 9.32. La valeur des paramètres a été obtenue par l'exploration d'une grille par validation croisée. Afin d'empêcher les variations dues au choix des échantillons, la validation est répétée 10 fois.

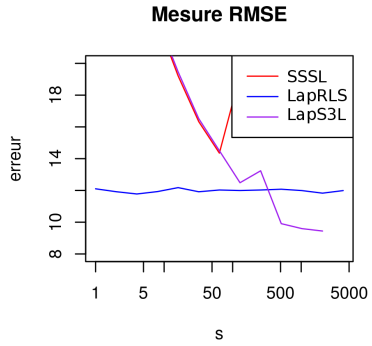

FIG. 1: Résultats sur la base de données Assainissement de la métropole de Lyon : Erreur RMSE (validation croisée) en fonction du nombre de valeurs propres

On peut remarquer que l'approche LapS3L affiche des résultats meilleurs que l'algorithme SSSL non régularisé, et meilleurs que la régularisation Laplacienne (LapRLS) pour la mesure d'erreur quadratique. Pour expliquer ces résultats, on peut regarder l'évolution de l'erreur en fonction de la valeur du paramètre s (figure 1). L'erreur pour la méthode SSSL trouve un minimum en fonction de s, puis diverge rapidement. Pour comprendre ce phénomène, il suffit de se rappeler que la deuxième étape de l'algorithme SSSL consiste à effectuer une régression aux moindres carrés ordinaires. Si s augmente considérablement, la régression tentera de classifier

des données correspondant à des valeurs propres trop faibles. Ces données sont assimilables à du bruit, et donc la méthode est victime de sur-apprentissage. La régularisation introduite dans la méthode proposée permet de mitiger ce risque dans une certaine mesure, raison pour laquelle la méthode proposée fournit de meilleurs résultats pour des valeurs de s plus grandes.

5 Conclusion

Nous avons proposé une approche de régression semi-supervisée qui combine deux approches intéressantes de l'état de l'art. La méthode proposée correspond aux exigences du cadre applicatif, et donne une erreur de régression quadratique plus faible que les deux méthodes de référence.

Cependant, la norme 2 employée dans la régularisation Laplacienne utilisée montre des limites dans le cas où le nombre de canalisations non datées augmente, alors que le nombre de canalisations datées reste constant [Alaoui et al. (2016)]. Bien que ce cas ne semble pas correspondre à notre problème, où il faut reconstituer les dates des canalisations *anciennes* uniquement, il serait intéressant de mettre en place une régularisation différente.

Remerciements

Ce travail a été réalisé grâce au soutien financier du LABEX IMU (ANR-10-LABX-0088) de l'Université de Lyon, dans le cadre du programme « Investissements d'Avenir » (ANR-11-IDEX-0007) géré par l'Agence Nationale de la Recherche (ANR).

Références

Ahmadi, M., F. Cherqui, J.-C. D. Massiac, et P. L. Gauffre (2014). Influence of available data on sewer inspection program efficiency. *Urban Water Journal 11*(8), 641–656.

Alaoui, A. E., X. Chen, A. Ramdas, M. J. Wainwright, et M. I. Jordan (2016). Asymptotic behavior of ℓ_p-based laplacian regularization in semi-supervised learning. In *COLT*, Volume 49 of *JMLR Workshop and Conference Proceedings*, pp. 879–906. JMLR.org.

Amini, M.-R. et P. Gallinari (2002). Semi-supervised logistic regression. In *Proceedings of the 15th European Conference on Artificial Intelligence*, pp. 390–394. IOS Press.

Azriel, D., L. D. Brown, M. Sklar, R. Berk, A. Buja, et L. Zhao (2016). Semi-supervised linear regression. *arXiv preprint arXiv :1612.02391*.

Basu, S., I. Davidson, et K. Wagstaff (2008). *Constrained clustering : Advances in algorithms, theory, and applications*. CRC Press.

Belkin, M., P. Niyogi, et V. Sindhwani (2006). Manifold regularization : A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research 7*, 2399–2434.

Bennett, K. P. et A. Demiriz (1999). Semi-supervised support vector machines. In *Advances in Neural Information processing systems*, pp. 368–374.

Blum, A. et S. Chawla (2001). Learning from labeled and unlabeled data using graph mincuts. In *ICML*, pp. 19–26. Morgan Kaufmann.

Blum, A. et T. Mitchell (1998). Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, pp. 92–100. ACM.

Cai, D., X. He, et J. Han (2006). Semi-supervised regression using spectral techniques. Technical report.

Chapelle, O., B. Schölkopf, et A. Zien (2006). *Semi-Supervised Learning*. MIT Press.

Harvey, R. R. et E. A. McBean (2014). Predicting the structural condition of individual sanitary sewer pipes with random forests. *Canadian Journal of Civil Engineering 41*(4), 294–303.

Ji, M., T. Yang, B. Lin, R. Jin, et J. Han (2012). A Simple Algorithm for Semi-supervised Learning with Improved Generalization Error Bound. *ICML 2012*.

Joachims, T. (1999). Transductive inference for text classification using support vector machines. In *ICML*, Volume 99, pp. 200–209.

Li, Y.-F., I. W. Tsang, J. T. Kwok, et Z.-H. Zhou (2013). Convex and scalable weakly labeled svms. *Journal of Machine Learning Research 14*, 2151–2188.

Mika, S., B. Schölkopf, A. J. Smola, K.-R. Müller, M. Scholz, et G. Rätsch (1999). Kernel PCA and de-noising in feature spaces. In *Advances in neural information processing systems*, pp. 536–542.

Moscovich, A., A. Jaffe, et B. Nadler (2016). Minimax-optimal semi-supervised regression on unknown manifolds. *arXiv preprint arXiv :1611.02221*.

Nigam, K., A. K. McCallum, S. Thrun, et T. Mitchell (2000). Text classification from labeled and unlabeled documents using em. *Machine learning 39*(2), 103–134.

Ryan, K. J. et M. V. Culp (2015). On semi-supervised linear regression in covariate shift problems. *Journal of Machine Learning Research 16*, 3183–3217.

Zhou, Z.-H. et M. Li (2005). Semi-supervised regression with co-training. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, IJCAI'05, San Francisco, CA, USA, pp. 908–913. Morgan Kaufmann Publishers Inc.

Zhu, X. (2006). Semi-supervised learning literature survey. *Computer Science, University of Wisconsin-Madison 2*(3), 4.

Summary

The installation date is often a primary consideration to explain the deterioration of the sanitation system. With this knowledge, the managers can (through deterioration models) predict the current condition of pipes that have not yet been examined, which is an essential information to take decision in a limited budget context. The data that has to be dealt with has different levels of complexity. The data has heterogeneous sources, a large volume, and limited information on its labelling (years): only 24% of the linear amount is known for the sanitation network. The underlying database will consist of the known features of the pipes (geometric profile, installation depth and so on). This study assessed the effect of some semi-supervised learning methods, and proposed a new approach suitable for this type of data.