Améliorer la classification semi-supervisée à base de graphes

Dino Ienco*, Ruggero G. Pensa**

*UMR TETIS, IRSTEA, Univ. Montpellier et LIRMM, Montpellier, France
dino.ienco@irstea.fr
**University of Turin, Computer Science Department, Turin, Italy
ruggero.pensa@unito.it

1 Introduction

La recherche sur l'apprentissage semi-supervisé (SSL) basé sur graphe (GBSSL) est principalement axée sur deux aspects : i) la construction du graphe des plus proches voisins et / ou ii) l'algorithme de propagation fournissant la classification. Nous nous intéressons dans ce poster à la représentation des données dans le but d'incorporer la semi-supervision au début du processus. Pour cela, nous apprenons un nouveau plongement de données à base de connaissance via un ensemble d'auto-encodeurs semi-supervisés pour améliorer les algorithmes GBSSL. La Figure 1(a) décrit le pipeline standard adopté pour le scénario GBSSL tandis que la Figure 1(b) décrit la stratégie proposée selon laquelle les informations d'étiquette sont exploitées plus tôt dans le pipeline. Notre contribution consiste à apprendre un plonge-

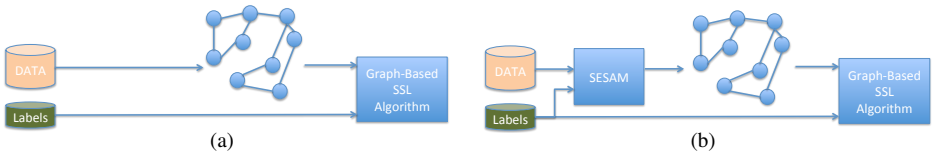

FIG. 1 – *Pipeline GBSSL standard (a) et SESAM (b).*

ment de données à base de connaissance qui alimente successivement un pipeline GBSSL standard. À cette fin, nous proposons d'utiliser un ensemble d'auto-encodeurs semi-supervisés (SSAE). La diversité est considérée comme une propriété clé dans la conception d'un schéma d'apprentissage d'ensemble (Chen et al., 2017). Pour chaque modèle de l'ensemble, nous échantillonnons de manière aléatoire la taille des différentes couches pour forcer la diversité. Nous appelons notre approche *SESAM*. Formellement, la fonctionne objective de chaque auto-encoder semi-supervsiée est définie comme suit : $L_{SSAE} = \frac{1}{|X|}\sum_{i=1}^{|X|}||X_i - AE(X_i)||^2 + \lambda - \frac{1}{|X^l|}\sum_{j=1}^{|X^l|}\sum_{c=1}^{|C|} y_{jc} \cdot log(\hat{y}_{jc})$ où X est l'ensemble de données, X^l l'ensemble de données étiquettées et c le numero de classes.

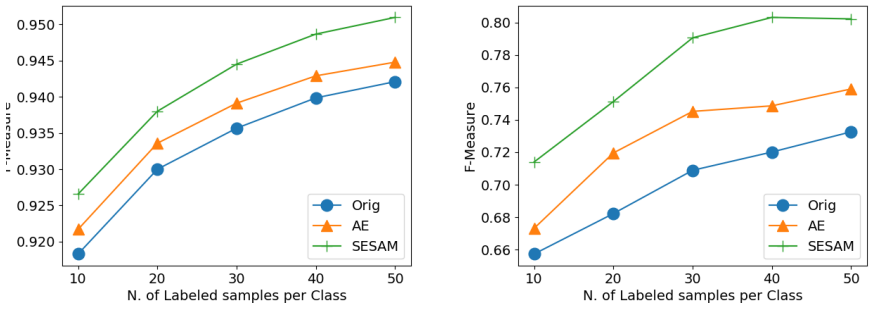

FIG. 2 – *F-Measure de l'algorithme CAMLP couplé avec les représentations* ORIG*,* AE *and SESAM en variant le nombre d'échantillons étiquetés par classe sur a) USPS and b) Sonar.*

2 Expériences

Nous comparons la représentation de données originale (*ORIG*) avec celle fournie par *SESAM* ainsi qu'une version non supervisée de *SESAM* qui empile les représentations induites par un ensemble d'auto-encodeurs entièrement non supervisés (*AE*). Les approches sont testés sur deux jeux de données : *Sonar* et *USPS*. Pour chaque représentation, nous construisons d'abord le graphe k-nearest neighbors mutuel (Suzuki and Hara, 2017) (K = 20) puis nous utilisons l'algorithme CAMLP (Yamaguchi et al., 2016) pour le classement final. Les performances sont analysées en fonction d'un niveau de supervision croissants, en faisant varier les exemples étiquetés, par classe, de 10 à 50. Nous répétons le processus de selection des échantillons 30 fois puis nous calculons la moyenne des résultats. Comme métrique d'évaluation, nous avons choisi la F-Measure. Nous définissons la taille de l'ensemble égale à 30. Pour entraîner le modèle, nous utilisons un taux d'apprentissage à 5×10^{-4} avec un facteur de décroissance de 5×10^{-5}. **Results** : La Figure 2 résume les résultats de la comparaison. Nous observons que la représentation fournie par *SESAM* fournit toujours les meilleurs résultats de classification indépendamment des valeurs des échantillons étiquetés par classe.

Références

J. Chen, S. Sathe, C. C. Aggarwal, and D. S. Turaga. 2017. Outlier Detection with Autoencoder Ensembles. In *SDM*. 90–98.

I. Suzuki and K. Hara. 2017. Centered kNN Graph for Semi-Supervised Learning. In *ACM SIGIR*. 857–860.

Y. Yamaguchi, C. Faloutsos, and H. Kitagawa. 2016. CAMLP : Confidence-Aware Modulated Label Propagation. In *SDM*. 513–521.