Visualisation de vecteurs supports – Stage M2 ou ingénieur

Romain Giot romain.giot@u-bordeaux.fr

Résumé du projet

Les Séparateurs à Vaste Marge (SVM) sont des classifieurs qui permettent de séparer deux classes en fonction d'un hyperplan séparateur (éventuellement non linéaire en utilisant une fonction noyau). Cet hyperplan est défini à l'aide de vecteurs supports appartenant à chacune des deux classes et disposé autour d'une marge. L'étape d'apprentissage du classifieur consiste à sélectionner un certain nombre de vecteur supports depuis le jeu de données d'apprentissage grâce à une étape d'optimisation contrainte par différents méta-paramètres (souplesse aux erreurs dans la construction de l'hyperplan, fonction noyau et ses paramètres). L'étape de classification consiste a effectuer une somme pondérée des appels du produit scalaire (ou de la fonction noyau) entre l'exemple à classifier et chaque vecteur support. Les performances de reconnaissance du SVM dépendent directement du choix de ces vecteurs supports (VS) censés bien représenter les données.

En fonction du jeu de données ou des méta-paramètres sélectionnés lors de l'apprentissage, les performances de reconnaissance peuvent être désastreuses. Il n'existe pas, à notre connaissance, de méthodes pour interpréter ces mauvaises performances en fonction de la visualisation des vecteurs supports. On peut dire la même chose pour la visualisation des vecteurs supports d'un système qui fonctionne convenablement.

Dans la littérature, nous pouvons trouver des méthodes pour visualiser les variables internes d'un réseau de neurone convolutionnel [3,4] ; ce qui montre que ce type de visualisation a donc un intérêt pour l'utilisateur d'outils d'apprentissage automatique (et donc de SVM).

Concernant la visualisation, le défi consiste à réussir à projeter en 2D des données de grandes dimension probablement non linéairement séparables de telle façon à rapidement identifier les vecteurs support des deux classes. D'autant plus que le paramètre de souplesse peut générer des vecteurs supports d'une classe plus proche des vecteurs de l'autre classe que de la sienne.

Travail a effectuer

Ce stage est orienté développement, les technologies à utiliser seront C++, OpenGl, Python.

Le but du stage est d'utiliser une implémentation existante de SVM (libSVM [1]) pour classer différents jeux de données de l'art puis récupérer leur performance de classification ainsi que les vecteurs supports.

Le stagiaire devra implémenter différentes métaphores visuelles pour mettre en évidence le fait que les vecteurs supports soient de bonne qualité ou non. Cette visualisation peut être basée sur des graphes de proximité, t-SNE [2] qui est un algorithme de projection de données de dimension moyenne (<50) en 2d, ou n'importe quelle autre information.

Une fois la ou les métaphores implémentées, il faudra les évaluer afin de vérifier que dans les jeux de données où le SVM fonctionne convenablement, les vecteurs supports de deux classes soient bien séparés. Tandis que dans les jeux de données plus difficiles la visualisation mette en évidence les vecteurs supports problématiques.

L'application a développer illustrera le fonctionnement du système (sélection d'un jeu de données et des paramètres, affichage des métriques de performance usuelles, affichage des vecteurs support).

Bibliographie

- [1] Chang, Chih-Chung, and Chih-Jen Lin. "LIBSVM: a library for support vector machines." ACM Transactions on Intelligent Systems and Technology (TIST) 2.3 (2011): 27.
- [2] Maaten, Laurens van der, and Geoffrey Hinton. "Visualizing data using t-SNE." Journal of Machine Learning Research 9.Nov (2008): 2579-2605.
- [3] Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034.
- [4] Zeiler, Matthew D., and Rob Fergus. "Visualizing and understanding convolutional networks." European Conference on Computer Vision. Springer International Publishing, 2014.