
Sparsification of Linear Models for Large-Scale Text Classification

Simon Moura^{*1}, Ioannis Partalas^{†2}, and Massih-Reza Amini^{‡1}

¹Laboratoire d'Informatique de Grenoble (LIG) – CNRS : UMR5217, Université Pierre-Mendès-France - Grenoble II, Institut polytechnique de Grenoble (Grenoble INP), Université Joseph Fourier - Grenoble I – UMR 5217 - Laboratoire LIG - 38041 Grenoble cedex 9 - France Tél. : +33 (0)4 76 51 43 61 - Fax : +33 (0)4 76 51 49 85, France

²WISEO - Objet Direct – VISEO - Objet Direct – Le Pulsar 4 avenue du Doyen Louis Weil 38000 GRENOBLE, France

Résumé

In this paper we propose a simple yet effective method for sparsifying a posteriori linear models for large-scale text classification. The objective is to maintain high performance while reducing the prediction time by producing very sparse models. This is especially important in real-case scenarios where one deploys predictive models in several machines across the network and constraints apply on the prediction time.

We empirically evaluate the proposed approach in a large collection of documents from the Large-Scale Hierarchical Text Classification Challenge. The comparison with a feature selection method and LASSO regularization shows that we achieve to obtain a sparse representation improving in the same time the classification performance.

Mots-Clés: Sparsification, Large scale text classification, Linear Models, Feature selection

*Intervenant

†Auteur correspondant: ioannis.partalas@viseo.com

‡Auteur correspondant: Massih-Reza.Amini@imag.fr