TreeBoost: A Boosting Algorithm for Multi-label Hierarchical Text Categorization

Tiziano Fagni
ISTI-CNR
Italy

Abstract:

Hierarchical Text Categorization (HTC) is the task of generating usually by means of supervised learning algorithms) text classifiers that operate on hierarchically structured classification schemes. Not withstanding the fact that most large-size classification schemes for text have a hierarchical structure, so far the attention of text classification researchers has mostly focused on algorithms for "flat" classification, i.e. algorithms that operate on non-hierarchical classification schemes. These algorithms, once applied to a hierarchical classification problem, are not capable of taking advantage of the information inherent in the class hierarchy, and may thus be suboptimal, in terms of efficiency and/or effectiveness. In this work we propose TreeBoost, a recursive hierarchical algorithm which can use as base learner one of the several variants of AdaBoost, a very well-known family of boosting algorithms. In particular on this work we focus on TreeBoost.MH, a variant which uses the popular AdaBoost.MH as base learner, and on TreeBoost.MP, a variant which uses MPBoost (a more optimized version of AdaBoost.MH introduced in one of our previous works) as base learner. We first introduce both AdaBoost.MH and MPBoost by describing their characteristics and the main differences between these two learning algorithms. Next we illustrate the TreeBoost algorithm by describing how the algorithm works and by analyzing its computational cost both at training and testing time. Finally we present the results of experimenting TreeBoost on two standard HTC benchmarks, and we discuss how TreeBoost variants compare to the corresponding "flat" baseline in terms of effectiveness/efficiency.