World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:10, No:09, 2016

Scalable Learning of Tree-Based Models on Sparsely Representable Data

Authors : Fares Hedayatit, Arnauld Joly, Panagiotis Papadimitriou

Abstract : Many machine learning tasks such as text annotation usually require training over very big datasets, e.g., millions of web documents, that can be represented in a sparse input space. State-of the-art tree-based ensemble algorithms cannot scale to such datasets, since they include operations whose running time is a function of the input space size rather than a function of the non-zero input elements. In this paper, we propose an efficient splitting algorithm to leverage input sparsity within decision tree methods. Our algorithm improves training time over sparse datasets by more than two orders of magnitude and it has been incorporated in the current version of scikit-learn.org, the most popular open source Python machine learning library.

Keywords : big data, sparsely representable data, tree-based models, scalable learning
Conference Title : ICDE 2016 : International Conference on Data Engineering
Conference Location : Chicago, United States
Conference Dates : September 19-20, 2016

Open Science Index, Computer and Information Engineering Vol:10, No:09, 2016 publications.waset.org/abstracts/52853.pdf