XGBoost: A Scalable Tree Boosting System

The boosting algorithm is an ensemble meta-algorithm used to reduce bias and variance in supervised learning. The algorithm converts weak learners into strong ones. The paper describes XGBoost, a scalable end-to-end tree boosting system. This is a well-known technique used by data scientists to achieve state-of-the-art results on many algorithm techniques. A scalable tree boosting system can be built using insights from this paper on cache access pattern, data compression, and sharding.

XGBoost is an open source machine learning system for tree boosting. It is scalable in all scenarios. On a single machine, the system runs ten times faster than existing solutions; in distributed or memory limited settings, it is scalable to billions of examples. Scalability of XGBoost is driven by optimization of systems and algorithms.

The four main topics covered in this paper are,

* Developing and designing a highly scalable end-to-end tree boosting algorithm.
* The development of a theoretically justified weight quantile sketch for efficient proposal calculation.
* Develop a novel sparsity-aware algorithm for parallel tree learning.
* Developing a cache-aware block structure for out-of-core tree learning.

This paper concludes that an end-to-end tree boosting system requires the construction of scalable tree boosting systems, such as cache patterns, during the building of XGBoost.