Machine Learning Test conference

Contribution ID: **9** Type: **not specified**

Large-Scale Machine Learning at Twitter

The success of data-driven solutions to difficult problems, along with the dropping costs of storing and processing massive amounts of data, has led to growing interest in large-scale machine learning. This paper presents a case study of Twitter's integration of machine learning tools into its existing Hadoop-based, Pig-centric analytics platform. We begin with an overview of this platform, which handles \traditional" data warehousing and business intelligence tasks for the organization. The core of this work lies in recent Pig extensions to provide predictive analytics capabilities that incorporate machine learning, focused specifically on supervised classification. In particular, we have identified stochastic gradient descent techniques for online learning and ensemble methods as being highly amenable to scaling out to large amounts of data. In our deployed solution, common machine learning tasks such as data sampling, feature generation, training, and testing can be accomplished directly in Pig, via carefully crafted loaders, storage functions, and user-defined functions. This means that machine learning is just another Pig script, which allows seamless integration with existing infrastructure for data management, scheduling, and monitoring in a production environment, as well as access to rich libraries of user-defined functions and the materialized output of other scripts.

Author: Mr JIMMY, Lin (Twitter)

Presenter: Mr JIMMY, Lin (Twitter)