Hivemall: Hive scalable machine learning library

Makoto YUI *and Isao Kojima

Information Technology Research Institute,
National Institute of Advanced Industrial Science and Technology, Japan.

Interest in Hadoop has clearly been growing, as many see the platform as an essential tool for analyzing big data. Along with this trend, Apache Hive [1] that provides a SQL-like frontend to Hadoop has become an essential instrument for analyzing data in Hadoop distributed file system (HDFS). Hive is useful for machine learning pipelines/workflows, especially for feature engineering as shown in our example[1], because it can facilitate easy data summarization, ad-hoc queries, and parallel joins of multiple resources. It is a natural consequence that users want to run machine-learning tasks against the large pool of data stored on HDFS as represented by Apache Mahout [2]. Though Mahout is a great tool to run machine learning tasks against data stored in HDFS, we found that the current implementation of classifiers, except random forest, is apparently not getting the full benefit of parallel MapReduce data processing[2].

To accomplish scalable machine learning on Hive/Hadoop, we are developing Hivemall [3, 4] on Github as an open source project. Hivemall is a scalable machine learning library licensed under the LGPL 2.1, providing machine learning functionality as well as feature engineering functions through UDFs/UDAFs/UDTFs of Hive. We designed Hivemall scalable to the number of training instances as well as the number of training features and showed in [5] that our scheme has competitive classification performance and superior training speed compared with state-of-the-art scalable machine learning frameworks; 5 and 7.65 times faster than Vowpal Wabbit [6] and Bismarck [7], respectively, for a click-through-rate (CTR) prediction task of the KDDCup 2012 track 2 [8].

We consider that Hivemall [3, 4] is a useful library for data scientists and developers who are familiar with Hive/Hadoop and the increasing attention proves our point. Hivemall provides support for the state-of-the-art online classifiers, e.g., Confidence Weighted [9], Adaptive Regularization of Weight Vectors [10], and Soft Confidence Weighted [11], and the development is moving toward the coming release of the initial major version.

References

[1] Apache Hive. http://hive.apache.org/.

[2] Apache Mahout. http://mahout.apache.org/.

[3] Hivemall project. https://github.com/myui/hivemall.

[4] Project details of Hivemall on mloss.org. http://mloss.org/software/view/510/.

[5] Makoto Yui and Isao Kojima. A Database-Hadoop Hybrid Approach to Scalable Machine Learning. In *Proc. IEEE 2nd International Congress on Big Data*, July 2013.

[6] Alekh Agarwal, Olivier Chapelle, Miroslav Dud'ık, and John Langford. A Reliable Effective Terascale Linear Learning System. *CoRR*, abs/1110.4198, 2011.

[7] Xixuan Feng, Arun Kumar, Benjamin Recht, and Christopher Ré. Towards a unified architecture for in-RDBMS analytics. In *Proc. SIGMOD*, pages 325–336, 2012.

[8] KDD Cup 2012, Track 2.
http://www.kddcup2012.org/c/kddcup2012-track2.

[9] Mark Dredze, Koby Crammer, and Fernando Pereira. Confidence-weighted linear classification. In *Proc. ICML*, pages 264–271, 2008.

[10] Koby Crammer, Alex Kulesza, and Mark Dredze. Adaptive regularization of weight vectors. *Machine Learning*, 91(2):155–187, 2013.

[11] Steven C. H. Hoi, Jialei Wang, and Peilin Zhao. Exact Soft Confidence-Weighted Learning. In *Proc. ICML*, 2012.

*m.yui@aist.go.jp

[1] https://github.com/myui/hivemall/wiki/KDDCup-2012-track-2-CTR-prediction-dataset

[2] You can find such examples in http://svn.apache.org/viewvc/mahout/trunk/examples/src/main/java/org/apache/mahout/classifier/sgd/