Building Segmented Models Using R and Hadoop

Collin Bennet[1], David Locke[1], Robert Grossman[1,2*], Steve Vecik[1]

1. Open Data Group
2. Laboratory for Advanced Computing, University of Illinois at Chicago
*Contact author: rlg1@opendatagroup.com

Keywords: segmented models, preprocessing data using Hadoop, scoring data using R and Hadoop

We introduce a simple framework for building and scoring models on datasets that span multiple disks in a cluster. We assume that through another analysis we know an appropriate means to segment the data. We use Hadoop to clean the data, preprocess the data, compute the derived features, and segment the data. We then invoke R individually on each of the segments and produce a model. By model here, we mean a model as formalized by Version 4.0 of the Predictive Model Markup Language (PMML) [1]; similarly, by segment, we mean a segment as formalized by the PMML Specification for Multiple Models, which includes both multiple models from ensembles and segments. We then gather the resulting models from each segment to produce a PMML multiple model file. Scoring is similar, except each node has access to the entire PMML multiple model file and not just the segment associated with the node. As in the first step that produces the multiple model file, preprocessing may use all the nodes, but each feature vector is sent to the appropriate node via segmentation for scoring. The framework is called Sawmill, depends upon Hadoop and R, and is open source. Sawmill also supports other parallel programming frameworks that generalize MapReduce, such as Sector/Sphere's User Defined Functions [2].

References

[1] PMML 4.0 - Multiple Models: Model Composition, Ensembles, and Segmentation, retrieved from www.dmg.org on February 10, 2010.

[2] Yunhong Gu and Robert L Grossman, Sector and Sphere: Towards Simplified Storage and Processing of Large Scale Distributed Data, Philosophical Transactions of the Royal Society A, Volume 367, Number 1897, pages 2429–2445, 2009. Sector/Sphere is open source and available from sector.sourceforge.net.