Term Project Proposal

Chao-Hong Chen, Qiuwei Shou

School of Informatics and Computing Indiana University Bloomington

The project aims to solve the multiclass linear regression classification on a large scale text dataset. The algorithm is used wildly in different fields to solve classification problem with discrete inputs. The traditional sequential approach shows the disadvantages of easily hitting the boundary of performance and scalability. Propose of the project is to adopt the traditional approach in a more open and distributive fashion. Collective library tools will be used to achieve parallel computation on MLR task. It brings up the potential to deal with more scalable dataset and speed up the performance.

Harp will be used to provide data abstraction and communication. It can be plugged into Hadoop runtime to achieve effective memory management, fault tolerance, hierarchal data abstraction and collective communication models. The parallel version of MLR will be developed on top of that. Open source code might be adopted from public sites, and if it's going to be used it will be specified and well cited in the project report.

Some research papers (Genkin, 2007) about MLR will be covered to help understand the algorithm and develop source code. Dataset preference is RCV1-V2 dataset and other datasets, OHSUMED or LSHTC, might be also used to compare the performance of algorithm on different datasets.

For the project responsibility, Qiuwei is main responsible for report and project design and Chao-Hong will be in charge of testing and developing, while both of the team members will work on the develop and test together. Project design is planned to be done by Oct 14. Development and test will starts from Oct 14 to Nov 21. A decent report will be delivered by Nov27 and task need to be complicated by Nov 25.

Reference：

[Genkin 2007] Genkin, A., Lewis, D.D., Madigan, D., 2007. Large-scale Bayesian logistic regression for text categorization. Technometrics 49, 291–304.