机器学习回归问题算法

在Hadoop环境下大规模并行实现

相关工作总结

在当今的信息化社会中，大规模数据集已经越来越普遍。特别是在机器学习和数据挖掘领域中，几百GB以至于TB的训练数据集就要求研究者开发效率更高、并行度更大的算法。而回归问题正是其中的一个典型代表。我们选择从罗吉斯回归模型[1]入手，因为这种模型除了解决回归问题，还可以用于二分类和多分类问题，其学习算法一般都具有良好的可扩展性。

在Hadoop环境下运用MapReduce实现机器学习算法的大规模并行，其他学者已经取得了不少优异的成果。早期的工作如谷歌公司实现的PSVM(*Parallel Support Vector Machines*)算法运用基于行的矩阵近似分解方法通过降低运算时内存大幅增加了程序并行度，成功将有效利用的机群节点数推进到上百台[2]。同样来自于谷歌公司的工作PLDA(*Parallel Latent Dirichlet Allocation*)通过有效地随机抽样方法同样提高了算法效率，并拥有对节点故障的鲁棒性[3]。最近由斯坦福大学Andrew Ng教授联合谷歌公司实现的针对深度学习算法的大规模并行化工作，不仅把并行机器数推进到上万，运算节点推进到上百万，更取得了目前最好的学习效果[4]。在针对回归问题的研究上，由卡纳基梅隆大学开发的GraphLab软件通过集成多种高效代码库和底层优化可以一定程度上提高算法效率[5]。但其设计初衷是在单机上利用多核实现并行，尽管后期发布了分布式版本，但仍不适合大规模多机并行。以上工作在Hadoop环境下均有开源代码提供，易于学习和研究。针对罗吉斯回归，我们已经从理论上提出了次线性的方法降低了算法复杂度[6]，并期望发展此算法，在大规模并行环境下更好得发挥出算法优势，取得优良的效果。

参考文献

【1】T. Hastie, R. Tishirani, and J. Friedman. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer-Verlag, New York, 2001

【2】Chang, E.Y., Zhu, K., Wang, H., Bai, H., Li, J., Qiu, Z., Cui, H.: Psvm: Parallelizing support vector machines on distributed computers. In: NIPS (2007)

【3】Yi Wang , Hongjie Bai , Matt Stanton , Wen-Yen Chen , Edward Y. Chang, PLDA: Parallel Latent Dirichlet Allocation for Large-Scale Applications, Proceedings of the 5th International Conference on Algorithmic Aspects in Information and Management, p.301-314, June 15-17, 2009, San Francisco, CA, USA

【4】Jeffrey Dean, Greg S Corrado, Rajat Monga, Kai Chen, Matthieu Devin, Quoc V Le, Mark Mao, Marc’Aurelio Ranzato, Andrew W Senior, Paul Tucker, Ke Yang, Andrew Y Ng, Large Scale Distributed Deep Networks. In: NIPS (2012)

【5】KYROLA, A., BLELLOCH, G., AND GUESTRIN, C. GraphChi: Large-scale graph computation on just a PC. In OSDI (2012)

【6】Haoruo Peng, Zhengyu Wang, Edward Y. Chang, Shuchang Zhou and Zhihua Zhang. Sublinear Algorithms for Penalized Logistic Regression in Massive Datasets. In Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD), 2012.