Parallel Data Mining

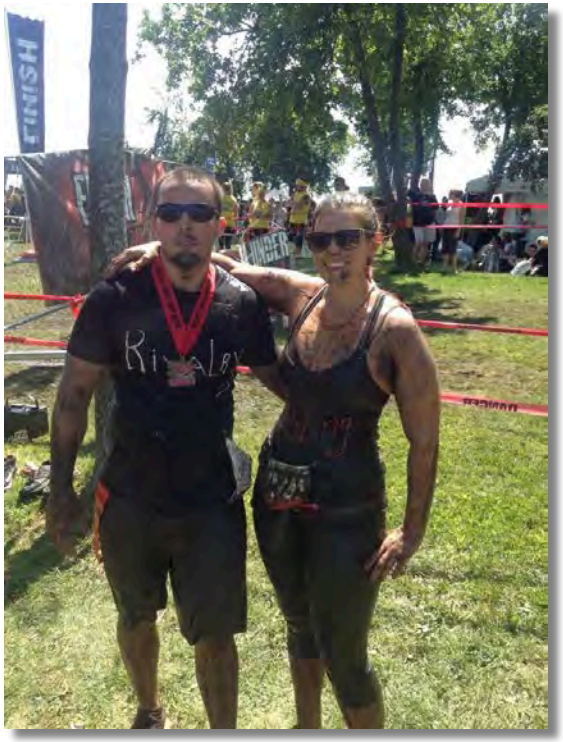
Team 2 – Flash Coders
Team Research Investigation
Presentation 1

Foundations of Parallel Computing
16 Sept 2014

Agenda

- Team Members
- Overview
- Computational Problem
- Sequential Algorithm
- Parallel Algorithm
- Reference Papers

Team Members



Overview

Big data analytics - data mining using parallel machine learning algorithms

Application dimensions:

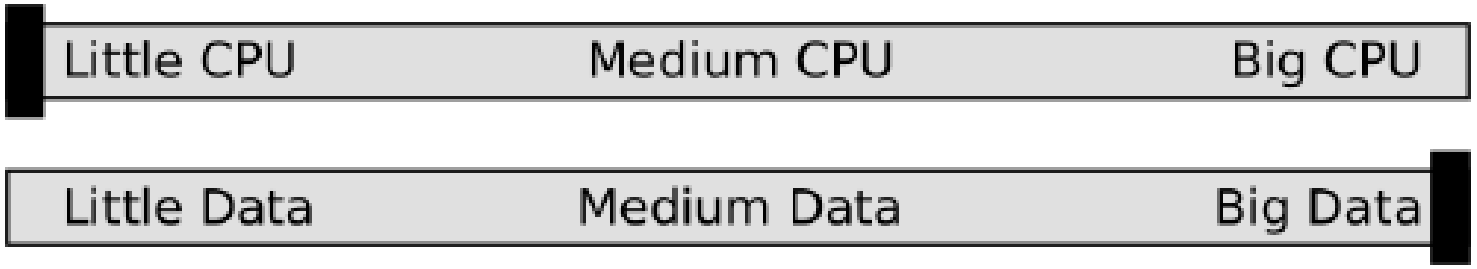
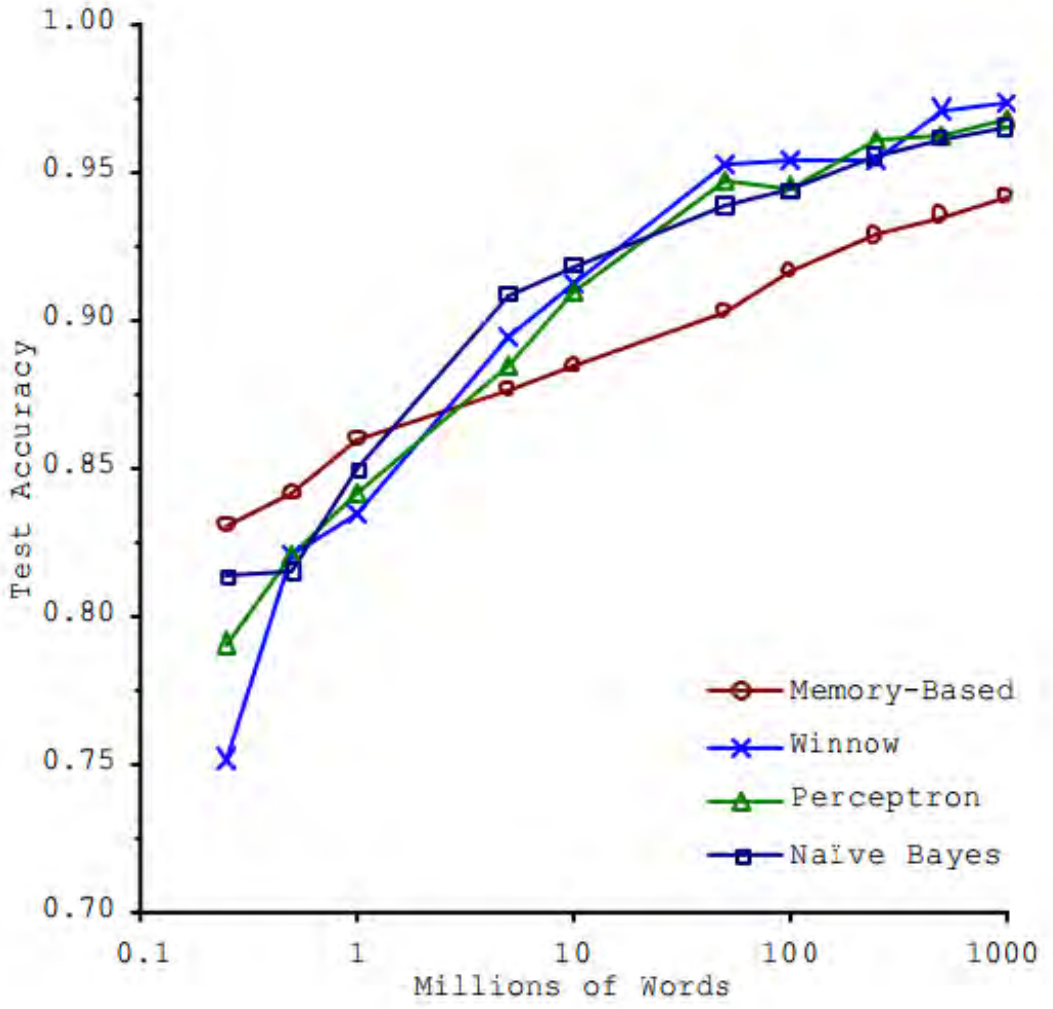


Figure from BIG CPU, BIG DATA: Solving the World's Toughest Computational Problems with Parallel Computing. Alan Kaminsky, 2014.

Why do we need Big Data?



- Data-not the algorithm used-is what matters
- Better to use all the data than to sample it

Figure from Banko and Brill, 2001, p2.

Where is it?

- Wal*Mart – 20 million point-of-sale transactions a day
- Facebook – 300 petabyte data warehouse that adds 600 terabytes a day
- YouTube – 4 billion views per day
- Instagram – 40 million photos per day

To learn from this data and others, we need to be able to analyze it faster than it arrives in order to make predictions & classifications in real time

Computational Problem

- Linear regression, logistic regression, & neural networks:
  - All try to minimize a cost function by adjusting certain parameters
  - All can be used for regression or classification problems
  - All require summations over the entire dataset which can take a very long time (in batch form)
  - All can get stuck in bad local optima and require several restarts to ensure good results

Computational Problem

Gradient descent can take a long time to converge with many training examples, and when it does it may not have found the global optimum

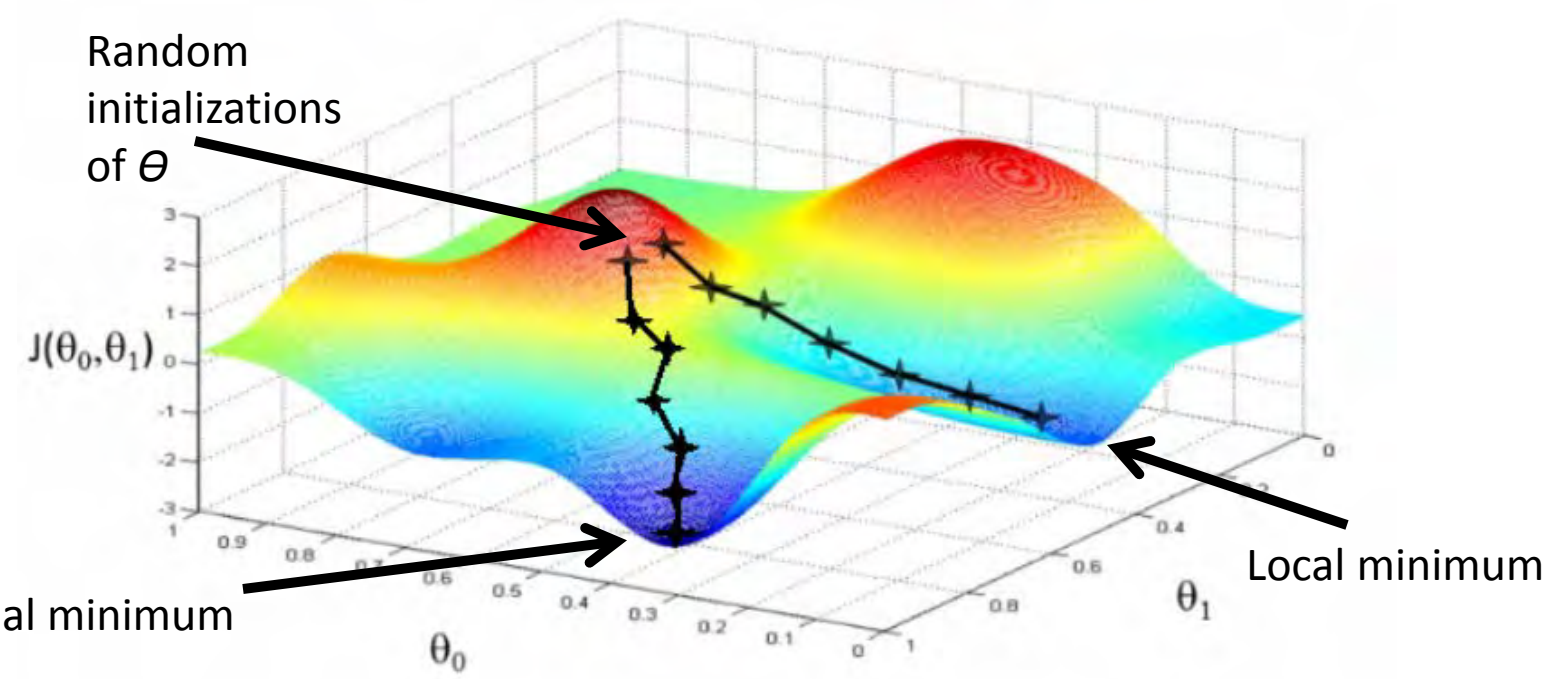


Figure from Vasilis Vryniotis. http://blog.datumbox.com/tuning-the-learning-rate-in-gradient-descent/. Accessed Sept 14, 2014.

Sequential Algorithm I

- Linear/logistic regression cost function:

$$\min_{\theta} J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^i) - y^i)^2$$

What if m equals 500 million training examples!

- Algorithm
  - Randomly initialize θ
  - Repeat until convergence (for all j=0,1,…,n):

$$\theta_j = \theta_j - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^i) - y^i) x_j{}^i$$

This term is $\frac{d}{d\theta_j} J(\theta)$

Sequential Algorithm II

- Predicted output is then $h_\theta(x^i)$ with trained parameters θ
  - Linear Regression:

$$h_\theta(x^i) = \sum_{j=1}^{n} \theta^i x^i$$

  - Logistic Regression

$$z = \sum_{j=1}^{n} \theta^i x^i$$

$$h_\theta(x^i) = {}^{1}/_{(1 + e^{-z})}$$

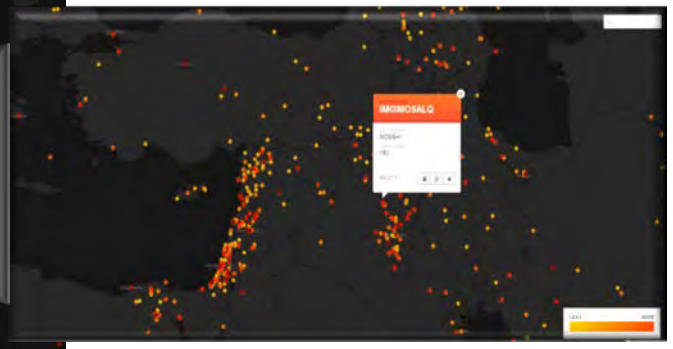
Parallel Algorithm

- Two ways to parallelize:
  - Within a cluster node: can use multiple cores to do the summation with a parallel reduction before each update to the parameters θ
  - Between cluster nodes: can use a different random initialization on each cluster node and take the best result as final solution

Dataset

- Global Data on Events, Location and Tone (GDELT) dataset contains over a quarter-billion records of global events from 1979 to the present
- Created with natural language processing of the daily articles of major international news sources
- Goldstein scale defines the violence and impact of an event on a scale from -10 to +10, which can be used to separate the most violent and destructive global events
- Recently made SQL accessible through Google BigQuery API
- URL: http://gdeltproject.org/

Visualizing GDELT



Zoomed in view of the Middle East with clickable events

Violent worldwide events: Nov 23-30, 2013

1st Citation

Mu Li, Tong Zhang, Yuqiang Chen, and Alexander J. Smola. 2014. Efficient mini-batch training for stochastic optimization. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining* (KDD '14). ACM, New York, NY, USA, 661-670.

URL: http://dl.acm.org/citation.cfm?id=2623330.2623612&coll=DL&dl=ACM&CFID=415399891&CFTOKEN=69514427

In this paper, the algorithm they present finds a more optimal convergence rate on large scale datasets by distributing the mini-batch (summation over a subset) training of the dataset. Both serial and distributed approaches show improved efficiency of convergence using this technique.

2nd Citation

Haoruo Peng, Ding Liang, and C. Choi. Evaluating parallel logistic regression models. In *Proceedings of the 2013 IEEE International Conference on Big Data*, pp 119-126, 6-9 Oct 2013.
URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6691743&isnumber=6690588

Logistic regression is expensive in terms of both time and storage for large datasets. This paper explores platform and algorithmic optimizations for LR training. It uses distributed platforms like Hadoop and paradigms like MapReduce to parallelize and make LR training on large data sets more efficient.

3rd Citation

Sameer Singh, Jeremy Kubica, Scott Larsen and Daria Sorokina. 2009. Parallel Large Scale Feature Selection for Logistic Regression. In *Proceedings of 2009 Society for Industrial and Applied Mathematics (SIAM) Data Mining*, SIAM, Philadelphia, PA, USA, 1172-1183.
URL: http://epubs.siam.org/doi/pdf/10.1137/1.9781611972795.100

In this paper, the authors have presented a parallel algorithm based on a MapReduce framework that evaluates the model's input features. They present a new heuristic and a parallel algorithm for feature evaluation which results in better performance and reduced computational cost.

Further References

- Michele Banko and Eric Brill. 2001. Scaling to very very large corpora for natural language disambiguation. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics* (ACL '01). Association for Computational Linguistics, Stroudsburg, PA, USA, 26-33.
  - URL: http://dl.acm.org/citation.cfm?id=1073017
- Mohammed Javeed Zaki. 1999. Parallel and Distributed Data Mining: An Introduction. In *Revised Papers from Large-Scale Parallel Data Mining, Workshop on Large-Scale Parallel KDD Systems, SIGKDD*, Mohammed Javeed Zaki and Ching-Tien Ho (Eds.). Springer-Verlag, London, UK, UK, 1-23.
  - URL: http://dl.acm.org/citation.cfm?id=744383
- Andrew Ng. Machine Learning course materials, Coursera. https://www.coursera.org/course/ml. Accessed September 10, 2014.

QUESTIONS