A Distributed Kernel Summation Framework for General Dimension Machine Learning

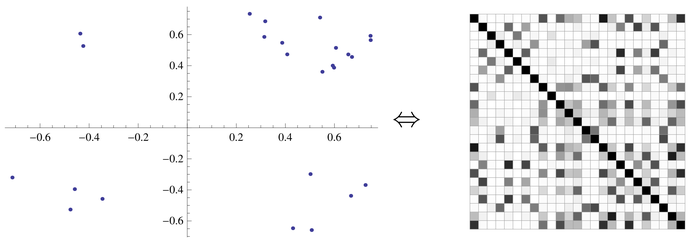
Best Paper Award at SIAM International Conference on Data Mining 2012 (Anaheim, CA)

Dongryeol Lee (drselee@gmail.com), Richard Vuduc (richie@cc.gatech.edu), Alexander G. Gray (agray@cc.gatech.edu)
Georgia Institute of Technology

Kernel Summations

Given two D-dimensional point sets $\mathbf{Q}$ and $\mathbf{R}$, compute $\forall \mathbf{q} \in \mathbf{Q}, \Phi(\mathbf{q}) = \sum_{\mathbf{r_j} \in \mathbf{R}} w_j k(\mathbf{q}, \mathbf{r_j})$ where $w_j \in \mathbb{R}$.

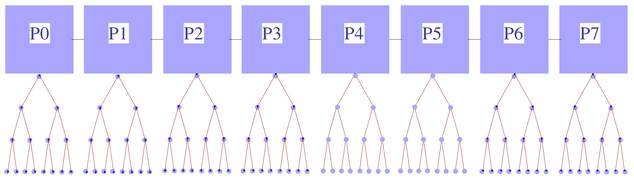
- The computational bottleneck ubiquitous in many machine learning methods.
- A kernel function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ defines a similarity measure between a pair of objects.



Example: Gaussian $\{K_{i,j}\}_{1 \le i,j \le N} = e^{-\|x_i - x_j\|^2/(2h^2)}$
Kernel summations computes approximately average similarity.

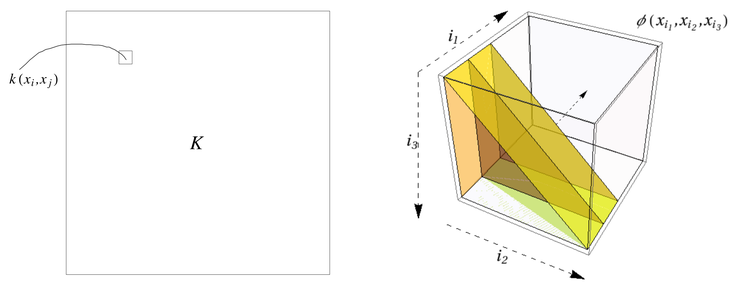
Distributed Data

If the disk space is cheap, why can't we store everything on one machine?



- More cost-effective to distribute data on a network of less powerful nodes than storing everything on one powerful node.
- Allows distributed query processing for high scalability.
- In some cases, all of the data cannot be stored on one node due to privacy concerns.

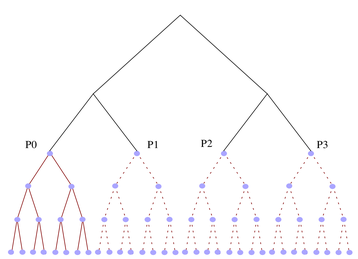
Scaling Kernel Methods

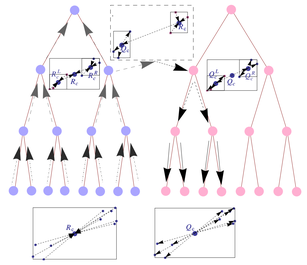


The computational cost seems super-quadratic in the number of points.

Contributions

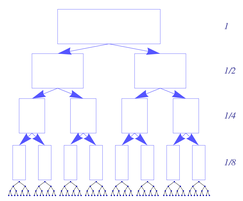
Parallel dual-tree methods using a distributed multidimensional tree such as kd-trees or metric trees. Uses distributed and shared memory parallelism using standards such as MPI, OpenMP, Intel TBB from the ground-up.





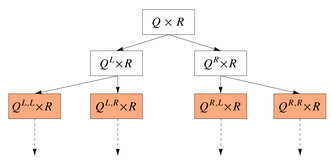
Parallel Indexing of Multidimensional Binary Trees

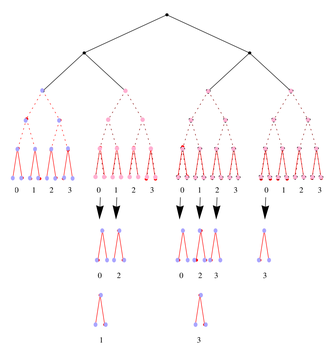
$\log p$ rounds of shuffling for building the distributed tree. Utilizes both distributed memory (MPI) and shared memory (OpenMP/Intel TBB) parallelism.



Parallel Computation Using Distributed Tree

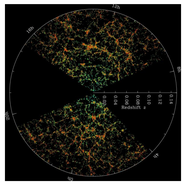
Pre-divide and spawn off independent computations. Each query subtree grabs necessary reference data.





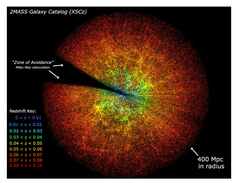
Kernel Summations in Scientific Applications

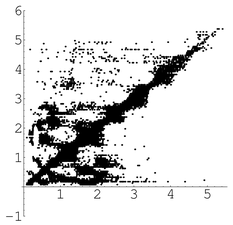
Sloan Digital Sky Survey collects around 200 GB data per day.

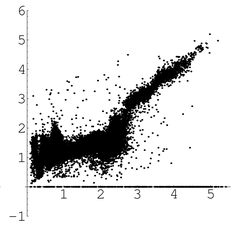


Kernel density estimator: $\max_{q \in Q} \sum_{r_j \in R} w_j k(q, r_j)$
Nadaraya-Watson: $\max_{q \in Q} \frac{\sum_{(r_j, y_j) \in R} w_j y_j k(q, r_j)}{\sum_{r_j \in R} w_j k(q, r_j)}$
Gaussian process regression: $K^{-1}y$

Large-scale redshift prediction of galaxies and quasars.

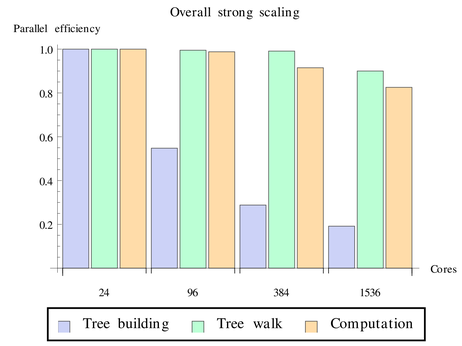






Experimental Results

Strong scaling results on a 10 million/4-dimensional subset of the SDSS dataset. Computed kernel density estimates using the Epanechnikov kernel with $h = 0.000030518$ (chosen by the plug-in rule) and $\epsilon = 0.1$. Raw numbers in seconds: (13.52, 339.36, 2371), (7.41, 24.38, 244), (2.93, 2.78, 98.78), (1.10, 0.27, 39.51)



Future Work

- Distributed computation on unreliable network connections.
- GPGPU-based acceleration.
- Parallelization of higher-order interactions including n-point correlation.