Efficient Data Reduction and Summarization

Ping Li

Department of Statistical Science

Faculty of Computing and Information Science

Cornell University

Two Contributions:

1. Data stream computations and entropy estimation

   Reduced sample size from 10^{10} or 10^7 to merely 10.

2. Hashing algorithms for large-scale high-dim search and learning

   (i) b-bit minwise hashing, (ii) one permutation hashing

   Reduced storage by 20-fold, dimensionality by 2^{20}-fold,

   and preprocessing cost by 500-fold.

Technical Approaches

1. **Data Stream Computations**: At time t, an entry of A_t is updated by (i_t, I_t): $\boxed{A_t[i_t] = A_{t-1}[i_t] + I_t}$. The task is to compute $F_{(\alpha)} = \sum_{i=1}^{D} |A_t[i]|^\alpha$, fast and and using small memory, which is crucial for anomaly detection.

   Our solution: Maximally-skewed stable random projections: $x = A_t R$.
   Entries of $R \in \mathbb{R}^{D \times k}$ are sampled from i.i.d. skewed α-stable distribution.
   Prior studies all used symmetric projections.

2. **BigData Search and Learning**: For a giant binary data matrix $A \in \mathbb{R}^{n \times D}$, the goals are (i) reducing # nonzeros; (ii) reducing dimensionality; (iii) highly efficient linear learning algorithms; (iv) sub-linear time near-neighbor search.

   Our solution: Permute the columns, store first nonzero location for each row using only the lowest b bits, break the space into k bins after one permutation.

Technical Achievements

1. **Skewed Projections for Nonnegative Data Streams**: If $A_t[i] \geq 0$, then $F_{(1)} = \sum_{i=1}^{D} |A_t[i]| = \sum_{i=1}^{D} A_t[i] = \sum_{s=0}^{t} I_s$ can be computed error-free. By " continuity", $F_{(\alpha)} = \sum_{i=1}^{D} A_t[i]^\alpha$ "should" also be easy.

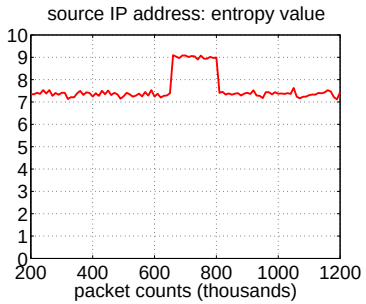
   Prior work using symmetric projections did not take advantage of this intuition.

   With skewed projections, $F_{(\alpha)}$ is estimated with variance $\propto O\left(|1-\alpha|^2\right)$. This easily solves the entropy estimation problem because, as $\alpha \to 1$, $\frac{1}{\alpha-1}\left(1 - \frac{F_{(\alpha)}}{F_{(1)}^\alpha}\right) \to$ Entropy. Otherwise, the variance blows up like $\frac{1}{|1-\alpha|^2}$.

2. **BigData Search and Learning**: Minwise hashing is the standard technique used in the search industry. Our method reduced the storage 20-fold, the dimensionality by 2^{20}-fold, and the preprocessing cost by 500-fold. We also develop novel methods for training large-scale learning and search.

Skewed Projections for Data Streams

Data stream: A very long vector of D items: $A_t[1]$, $A_t[2]$, ..., $A_t[D]$. $D \geq 2^{64}$ in many network applications. A DDoS attack changes the distribution of traffic, which could be captured by **Shannon entropy** measurements.



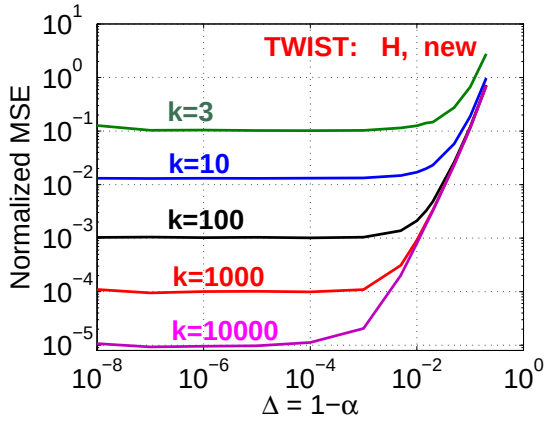
Maximally-skewed stable projections: $x_j = [A_t \times R]_j$, $j = 1$ to k. Entries of R are sampled from skewed stable distributions.

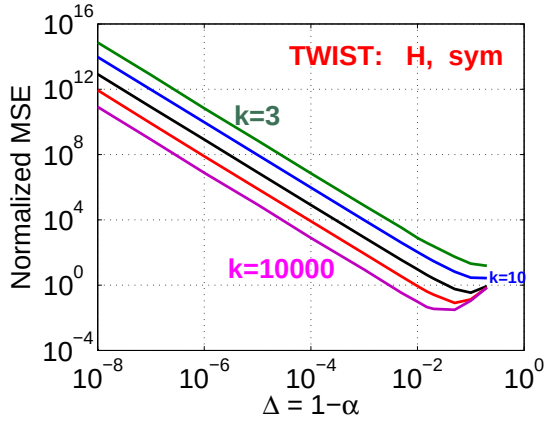
$$\hat{F}_{(\alpha),new} = \frac{1}{\Delta^\Delta}\left[\frac{k}{\sum_{j=1}^{k} x_j^{-\alpha/\Delta}}\right]^\Delta, \quad \Delta = 1-\alpha$$

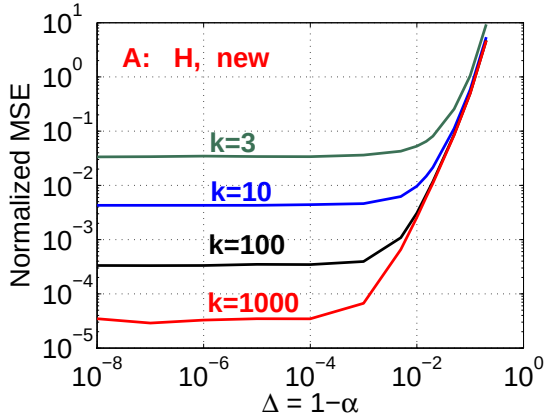
$$Var\left(\hat{F}_{(\alpha),new}\right) = \Delta^2 \frac{F_{(\alpha)}^2}{k}\left((3-2\Delta) + O\left(\frac{1}{k}\right)\right).$$

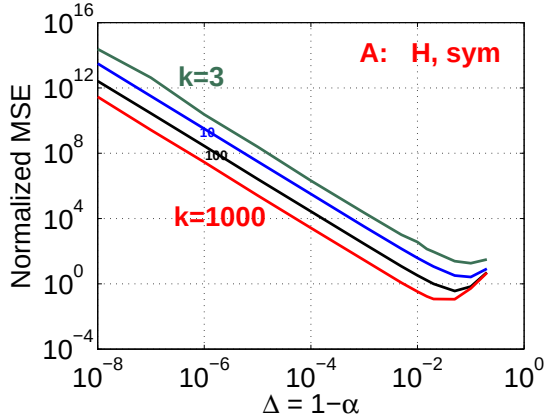
Entropy Estimation Using Symmetric & Skewed Projections

Data: Freq. of word occurrences in docs. Heavy-tailed similar to network data.







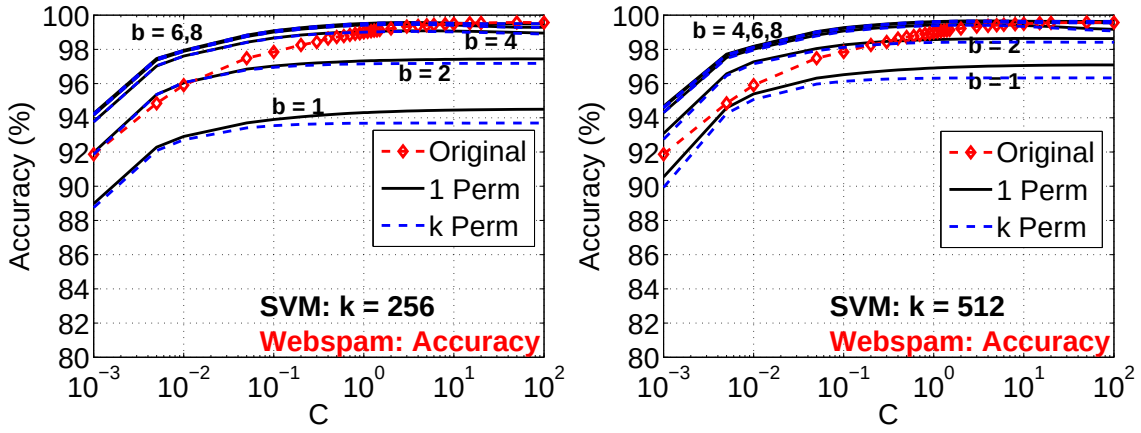


One Permutation Hashing

Example: Apply a permutation π on the columns, divide the space evenly into $k = 4$ bins and select the smallest nonzero in each bin, using only b bits.

	1				2				3				4			
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
$\pi(S_1)$:	0	0	1	0	1	0	0	1	0	0	0	0	0	1	0	0
$\pi(S_2)$:	1	0	0	1	0	0	1	0	0	0	0	0	0	1	0	0
$\pi(S_3)$:	1	1	0	0	0	0	0	0	0	0	1	0	1	0	0	0

Linear SVM experiments on Webspam data: one permutation is even better.



Impacts of Results

1. **Compressed Counting for Data Stream Computations**: Entropy estimation is a known difficult problem. Skewed projections made it trivial.

   Applications of data stream computations are mainly in the network area. Engineers and researchers have approached the PI regarding the implementations for network anomaly detection and router design.

2. **Hashing Algorithms for Large-scale Search and Learning:** Many real-world applications may benefit from these hashing algorithms. An earlier work was featured by the Commun. of the ACM in 2011. As far as the PI knows, companies are already using b-bit minwise hashing and its applications in search and linear learning. Even the most recent work on one permutation hashing has already been implemented.

Extensions and Outreach

Other related funding: NSF#1249316 EAGER: *Preliminary Study of Hashing Algorithms for Large-Scale Learning*, 09/2012 - 08/2013, $100K.

Research highlights: Li and König, *Theory and Applications of b-Bit Minwise Hashing*, Communications of the ACM, August 2011.

Yahoo! Learning to Rank Grand Challenge: using own boosting algorithms.

Conferences attended: MMDS, NVAC, VAST, SC, KDD, WWW, NIPS.

Industry presentations: MSR-Redmond, IBM-TJW, Google-NYC, Google-MV.