Learning and Data Selection in Big Datasets

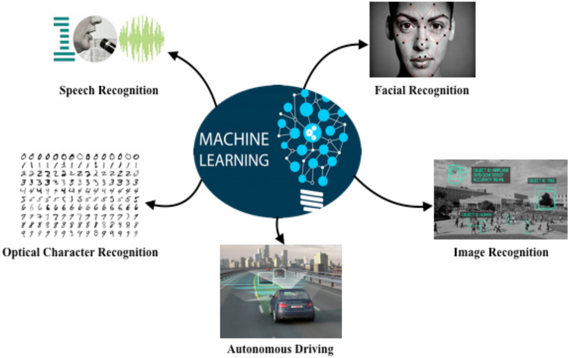
H. S. Ghadikolaei, H. Ghauch, C. Fischione, and M. Skoglund

School of Electrical Engineering and Computer Science
KTH Royal Institute of Technology
Stockholm, Sweden

http://www.kth.se/profile/hshokri
hshokri@kth.se

International Conference on Machine Learning (ICML)
Long Beach, CA, USA, June 2019

Big data era



- Outstanding performance of ML
  - Usually trained over massive datasets
  - Examples: MNIST (70k samples) and MovieLens (20M samples)

What about a small set of critical samples that best describes an unknown model?

Related works

- Experiment design [Sacks-Welch-Mitchell-Wynn, 1989]
  - to minimize total labeling cost
  - different setting
- Active learning [Settles, 2012]
  - to minimize total labeling cost
  - different setting
- Core set selection [Tsang-Kwok-Cheung, 2005]
  - to find a small representative dataset
  - limited to SVM
- Influence score [Koh-Liang, 2017]
  - to understand the importance of every sample
  - greedy: cannot score a set of samples

Our approach

Conventional training: (ℓ_i: loss of sample i, N: dataset size, h: parameterized function from space $\mathcal{H}$)

$$\underset{h\in\mathcal{H}}{\text{minimize}} \ \frac{1}{N}\sum_{i=1}^{N}\ell_i(h) \ .$$

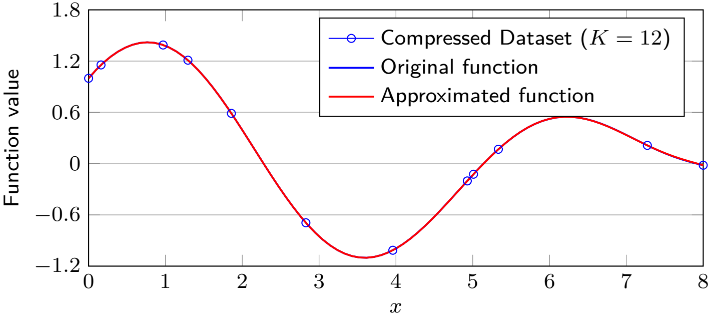
Our proposal: (joint learning and data selection)

$$\underset{h\in\mathcal{H},\mathbf{z}\in\{0,1\}^N}{\text{minimize}} \quad \frac{1}{\mathbf{1}^T\mathbf{z}}\sum_{i=1}^{N} z_i\ell_i(h), \quad \text{s.t.} \ \frac{1}{N}\sum_{i=1}^{N}\ell_i(h) \leq \epsilon \ , \ \mathbf{1}^T\mathbf{z} \geq K \ .$$

- Maximum compression rate: $1 - K/N$
- Solved efficiently using our proposed Alternating Data Selection and Function Approximation algorithm
- Under some regularity assumptions, $K \geq \lceil (1 + 2LT\sqrt{d/\delta})^d \rceil$ samples are enough for learning an L-Lipschitz function defined on interval $[0,T]^d$ with arbitrary accuracy δ ($\delta \leq \epsilon$)

Experimental results

Illustrative example:



Real-world data sets (from UCI repos.):

- experiments on Individual household electric power consumption ($N = 1.5M$, $d = 9$) and YearPredictionMSD ($N = 463K$, $d = 90$) datasets
- almost no loss in learning performance after **95% compression** using our approach

Final remarks

- Theoretically, almost 100% compressibility of big data is feasible without a noticeable drop in the learning performance
- Much faster training over the small representative dataset
- Inefficiency of the existing approaches to create datasets (which lead to a massive amounts of redundancy)
- **Applications:**
  - edge computing: reducing the communication overhead
  - IoT: enabling low-latency learning and inference over a communication-limited network

Visit our poster: Pacific Ballroom #170

References

- J. Sacks, W.J. Welch, T.J. Mitchell, and H.P. Wynn, "Design and analysis of computer experiments," *Statistical Science*, 1989.
- B. Settles, "Active learning," *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 2012.
- I.W. Tsang, J.T. Kwok, and P.M. Cheung, "Core vector machines: Fast SVM training on very large data sets," *Journal of Machine Learning Research*, 2005.
- P.W. Koh, and P. Liang, "Understanding black-box predictions via influence functions," in *Proc. International Conference on Machine Learning*, 2017.

ROYAL INSTITUTE
OF TECHNOLOGY

Learning and Data Selection in Big Datasets

H. S. Ghadikolaei, H. Ghauch, C. Fischione, and M. Skoglund

School of Electrical Engineering and Computer Science
KTH Royal Institute of Technology
Stockholm, Sweden

http://www.kth.se/profile/hshokri
hshokri@kth.se

International Conference on Machine Learning (ICML)
Long Beach, CA, USA, June 2019