Learning and Data Selection in Big Datasets

Hossein S. Ghadikolaei [1] Hadi Ghauch [1 2]
Carlo Fischione [1] Mikael Skoglund 1

[1]School of Electrical Engineering and Computer Science,
KTH Royal Institute of Technology, Stockholm, Sweden
[2]COMELEC Department, Telecom ParisTech, Paris, France

ICML-2019

Introduction

- Background: state-of-the-art machine learning methods often need to be trained on increasingly large datasets
- Goal: use a small dataset to train a good model

Core-set selection

- Different from active learning--reducing the total labeling cost
- it focuses on finding a small representative set of samples with cardinality K in a big dataset of labeled samples with cardinality N

Problem setting

$$\underset{h\in\mathcal{F},\boldsymbol{z}}{\arg\min}\ g(h,\boldsymbol{z}) := \frac{1}{\mathbf{1}^T\boldsymbol{z}}\sum\nolimits_{i\in[N]} z_i\ell_i(h) \qquad (2a)$$

$$\text{s.t.}\ \ g_1(h) := \frac{1}{N}\sum\nolimits_{i\in[N]} \ell_i(h) \le \epsilon\,, \qquad (2b)$$

$$g_2(\boldsymbol{z}) := \mathbf{1}^T\boldsymbol{z} \ge K\,,\ \boldsymbol{z}\in\{0,1\}^N, \qquad (2c)$$

l(h): loss of single data
Z: binary vector with dimensionality N

2a: prevent overfitting
2b: prevent trivial solution(e.g. empty set)

Solution

$$(P2a): \quad \boldsymbol{z}^{(k+1)} \in \underset{\boldsymbol{z}\in\{0,1\}^N}{\arg\min}\ g(h^{(k)}, \boldsymbol{z}), \tag{3a}$$

$$\text{s.t. } g_2(\boldsymbol{z}) \geq K\,, \tag{3b}$$

D-step:
Select data

$$(P2b): \quad h^{(k+1)} \in \underset{h\in\mathcal{F}}{\arg\min}\ g(h, \boldsymbol{z}^{(k+1)}), \tag{4a}$$

$$\text{s.t. } g_1(h) \leq \epsilon\,. \tag{4b}$$

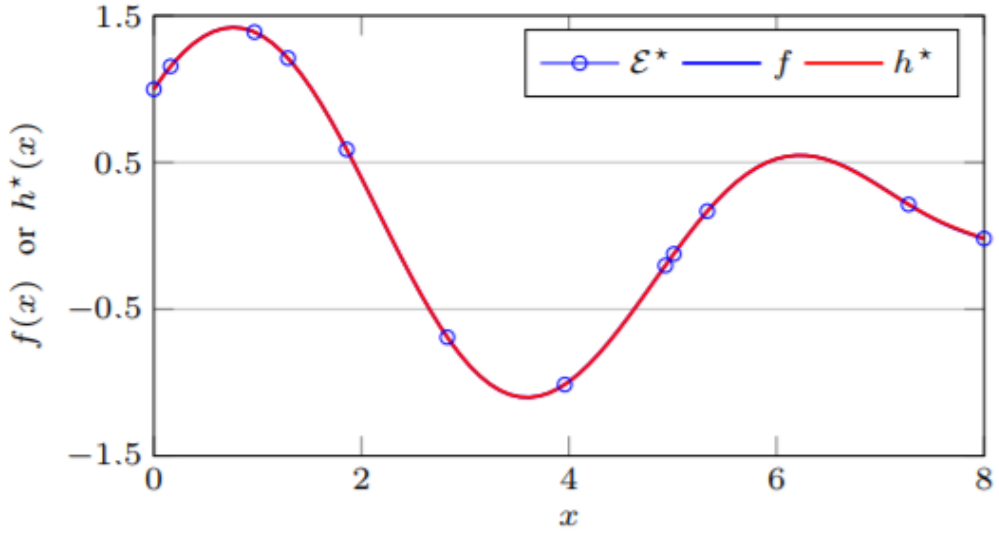
F-step:
Train model

Algorithm

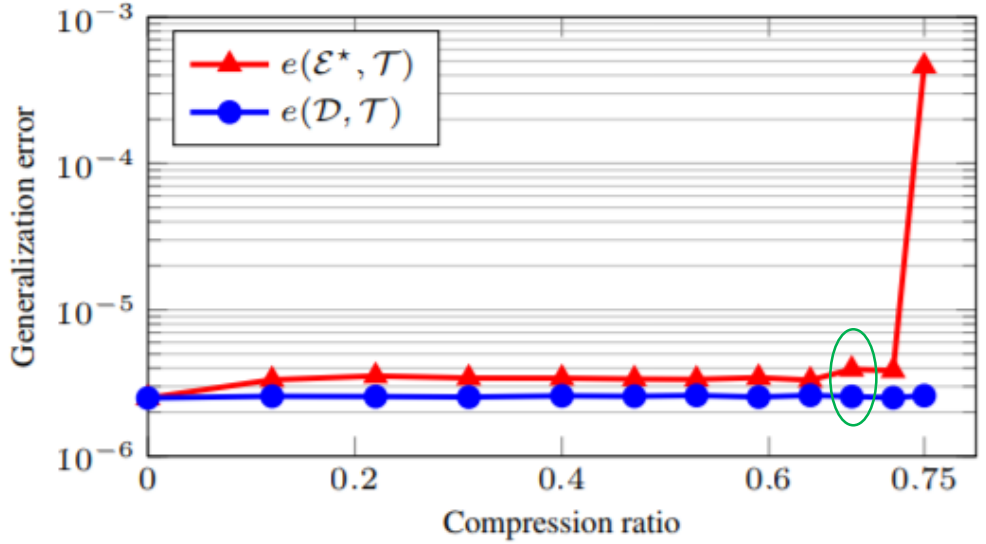
Algorithm 1 Alternating Data Selection and Function Approximation (DF)

Initialize: $z^{(1)} = \mathbf{1}$
for $k = 1, 2, 3, \ldots$ **do**
  // F-step
  Update $h^{(k+1)}$ by solving $(P2b)$ in (7)
  // D-step
  Compute $c_i^{(k)}$, $\forall i \in [N]$
  Update $z^{(k+1)}$ by solving $(P2a)$ in (5), using Proposition 1
  Break if $|g(h^{(k+1)}, z^{(k+1)}) - g(h^{(k)}, z^{(k)})| < \gamma$
end for
Return: $h^{(k+1)}$ and $z^{(k+1)}$

Experiment



(a) Function f and an example of $\mathcal{E}^\star$ with $K = 12$

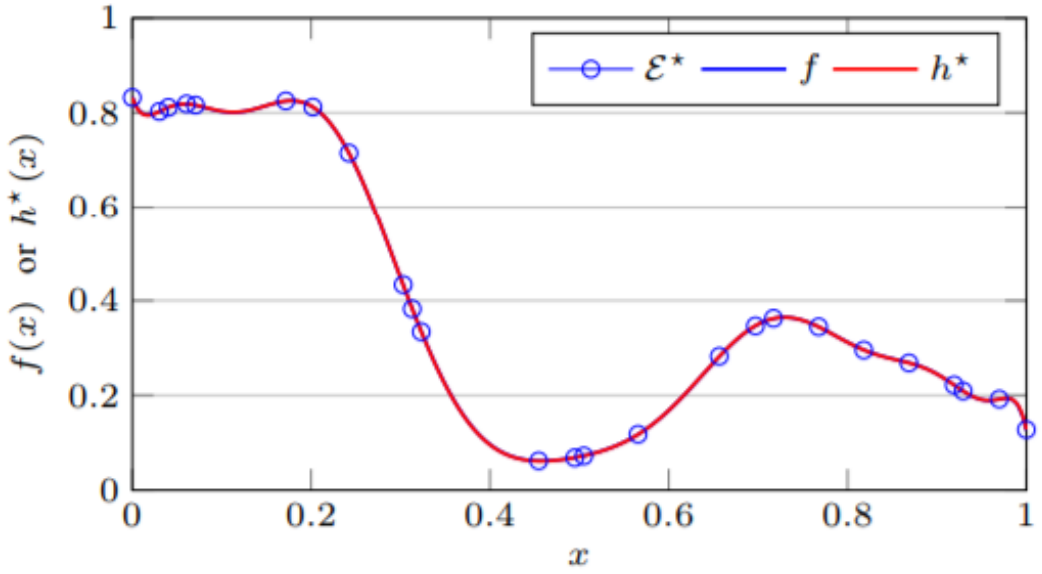


(b) Generalization error

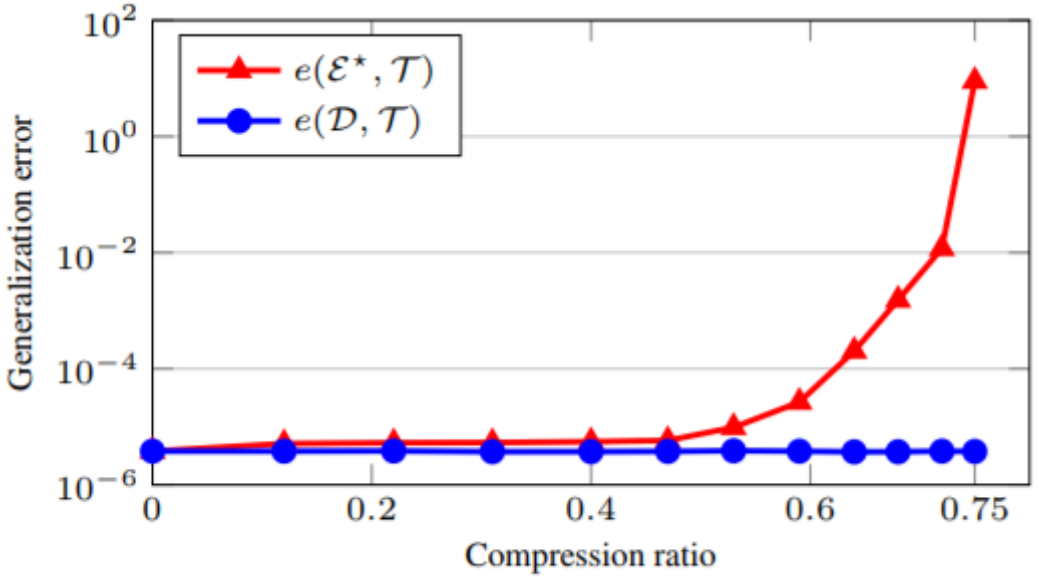
f: real function
h*: learned function
N=100

Compression ratio： 1-K/N

Experiment



(a) Function f and an example of $\mathcal{E}^\star$ with $K = 25$.



(b) Generalization error

Experiment

168

46W

CR	Algorithm	Bodyfat	Housing	Space-ga	YearPredictionMSD	Power Consumption
0%	Proposed	0.0245 ± 0.0051	0.0301 ± 0.0056	0.1323 ± 0.0134	0.0082 ± 0.0007	0.0142 ± 0.0008
25%	Proposed	0.0294 ± 0.0058	0.0345 ± 0.0071	0.1325 ± 0.0142	0.0083 ± 0.0007	0.0144 ± 0.0008
	Influence	0.0298 ± 0.0055	0.0345 ± 0.0077	0.1326 ± 0.0144	0.0083 ± 0.0007	0.0144 ± 0.0008
	Random	0.0315 ± 0.0081	0.0347 ± 0.0092	0.1329 ± 0.0145	0.0083 ± 0.0008	0.0145 ± 0.0008
50%	Proposed	0.0333 ± 0.0067	0.0348 ± 0.0080	0.1338 ± 0.0175	0.0085 ± 0.0008	0.0144 ± 0.0009
	Influence	0.0342 ± 0.0071	0.0349 ± 0.0078	0.1340 ± 0.0175	0.0087 ± 0.0008	0.0144 ± 0.0008
	Random	0.0370 ± 0.0102	0.0361 ± 0.0105	0.1412 ± 0.0202	0.0162 ± 0.0071	0.0190 ± 0.0124
75%	Ours	0.0360 ± 0.0060	0.0374 ± 0.0076	0.1351 ± 0.0169	0.0086 ± 0.0010	0.0145 ± 0.0008
	Influence	0.0383 ± 0.0067	0.0378 ± 0.0081	0.1519 ± 0.0231	0.0091 ± 0.0008	0.0146 ± 0.0008
	Random	0.0411 ± 0.0166	0.0392 ± 0.0099	0.1901 ± 0.0270	0.0312 ± 0.0156	0.0357 ± 0.0235
80%	Proposed	0.0417 ± 0.0077	0.0382 ± 0.0076	0.1354 ± 0.0171	0.0086 ± 0.0009	0.0145 ± 0.0008
	Influence	0.0446 ± 0.0097	0.0384 ± 0.0080	0.1535 ± 0.0240	0.0099 ± 0.0015	0.0151 ± 0.0011
	Random	0.0461 ± 0.0168	0.0395 ± 0.0107	0.2006 ± 0.0281	0.0370 ± 0.0171	0.0380 ± 0.0254
95%	Proposed	0.0630 ± 0.0134	0.0538 ± 0.0122	0.1386 ± 0.0217	0.0088 ± 0.0012	0.0153 ± 0.0011
	Influence	0.0753 ± 0.0101	0.0565 ± 0.0115	0.1951 ± 0.0329	0.0142 ± 0.0073	0.0216 ± 0.0009
	Random	0.0944 ± 0.0540	0.0744 ± 0.0372	0.3713 ± 0.0714	0.0560 ± 0.0478	0.0749 ± 0.0628

Database	# Training samples
Bodyfat	168
Housing	337
Space-ga	2,071
YearPredictionMSD	463,715
Power Consumption	1,556,445

Influence and random: Influence-based and random data selection