A Learning Augmented approach to Cardinality Estimation

Mărcuș Alexandru Marian

November 26, 2025

1 Schita cuprins

2 Theoretical Foundation

consistency, robustness, competitiveness

loss function: log loss/ cross entropy $L = (\log(E) - \log(N))^2$

3 Plan pentru partea aplicativa a lucrarii

Implementarea standard a algoritmelor HLL si HLL++.
Generare de date cu diverse distributii ce ar putea fi regasite in date reale (ex. uniform, clustered etc).
Feature extraction din HLL sketch (ex. max, min, media, devia standard, % din registre sunt goale, histograma a valorilor din registrii).
Antrenarea unui model mic (ex. regresie, un neural network cu putine layere etc).
Evaluarea modelului fata de standarde ca si acuratete, runtime si memorie.
Analiza rezultatelor si concluzii.

4 Methodology

4.1 Overview

We implement the standard HLL algorithm and augment it with a learned post-processing module. The neural component receives various features from the final register array M and uses them for deciding the weight of each register in the final result computation.

4.2 Data

4.3

Algorithm 1 Learned HyperLogLog

Require: Let $h : \mathcal{D} \to \{0,1\}^{64}$ hash data from domain $\mathcal{D}$. Let $m = 2^p$ with $p \in [4..16]$.

Phase 0: Initialization.

Define $\alpha_{16} = 0.673,\ \alpha_{32} = 0.697,\ \alpha_{64} = 0.709,$
  $\alpha_m = 0.7213/(1 + 1.079/m)$ for $m \geq 128$.
Initialize m registers $M[0]$ to $M[m-1]$ to 0.

Phase 1: Aggregation.

for all $v \in S$ **do**
  $x := h(v)$
  $id := (x_{63}, \ldots, x_{64-p})_2$ ▷ First p bits of x
  $w := (x_{63-p}, \ldots, x_0)_2$
  $M[id] := \max(M[id], \rho(w))$
end for

Phase 2: Result computation.

return $E := \alpha_m m^2 \left(\sum_{j=0}^{m-1} 2^{-M[j]} \right)^{-1}$ ▷ The "raw" estimate

References

[1] Xiaoying Wang, Changbo Qu, Weiyuan Wu, Jiannan Wang, and Qingqing Zhou. Are we ready for learned cardinality estimation? *CoRR*, abs/2012.06743, 2020.

[2] Stefan Heule, Marc Nunkesser, and Alexander Hall. Hyperloglog in practice: algorithmic engineering of a state of the art cardinality estimation algorithm. In *Proceedings of the 16th International Conference on Extending Database Technology*, EDBT '13, page 683–692, New York, NY, USA, 2013. Association for Computing Machinery.

[3] Tim Kraska, Alex Beutel, Ed H. Chi, Jeffrey Dean, and Neoklis Polyzotis. The case for learned index structures. *CoRR*, abs/1712.01208, 2017.

[4] Zhenwei Dai and Anshumali Shrivastava. Adaptive learned bloom filter (ada-bf): Efficient utilization of the classifier. *CoRR*, abs/1910.09131, 2019.

[5] Michael Mitzenmacher and Sergei Vassilvitskii. Algorithms with predictions. *CoRR*, abs/2006.09123, 2020.

[6] Philippe Flajolet, Éric Fusy, Olivier Gandouet, and Frédéric Meunier. Hyperloglog: the analysis of a near-optimal cardinality estimation algorithm. *Discrete Mathematics & Theoretical Computer Science*, DMTCS Proceedings vol. AH, 2007 Conference on Analysis of Algorithms (AofA 07), Jan 2007.

[7] Michael Mitzenmacher. A model for learned bloom filters and related structures. *CoRR*, abs/1802.00884, 2018.

[8] Kenneth G. Paterson and Mathilde Raynal. HyperLogLog: Exponentially bad in adversarial settings. Cryptology ePrint Archive, Paper 2021/1139, 2021.

[9] Qingzhi Ma, Ali Mohammadi Shanghooshabad, Mehrdad Almasi, Meghdad Kurmanji, and P. Triantafillou. Learned approximate query processing: Make it light, accurate and fast. In *Conference on Innovative Data Systems Research*, 2021.

[10] Thodoris Lykouris and Sergei Vassilvitskii. Competitive caching with machine learned advice. *CoRR*, abs/1802.05399, 2018.

[11] Keerti Anand, Rong Ge, and Debmalya Panigrahi. Customizing ml predictions for online algorithms, 2022.

[12] Otmar Ertl. New cardinality estimation algorithms for hyperloglog sketches. *CoRR*, abs/1702.01284, 2017.

[13] Kangfei Zhao, {Jeffrey Xu} Yu, Hao Zhang, Qiyan Li, and Yu Rong. A learned sketch for subgraph counting. *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 2142–2155, 2021. Publisher Copyright: © 2021 ACM.; 2021 International Conference on Management of Data, SIGMOD 2021 ; Conference date: 20-06-2021 Through 25-06-2021.

[14] Brian Tsan, Asoke Datta, Yesdaulet Izenov, and Florin Rusu. Approximate sketches. *Proc. ACM Manag. Data*, 2(1), March 2024.

[15] Aiyou Chen, Jin Cao, Larry Shepp, and Tuan Nguyen. Distinct counting with a self-learning bitmap. *Journal of American Statistical Association*, 106:879–890, 2011.