Partitioned Learned Count-Min Sketch

THUY TRANG NGUYEN* and CAMERON MUSCO*, University of Massachusetts Amherst, USA

We consider learning augmented algorithms for the fundamental problem of frequent item identification in data streams. The learned Count-Min sketch (LCMS) of Hsu et al. [1] combines a standard Count-Min sketch frequency estimation data structure with a learned model, by partitioning items in the input stream into two sets. Items with sufficiently high predicted frequencies have their frequencies tracked exactly, while the remaining items, with low predicted frequencies, are placed into the Count-Min sketch data structure. Following an approach introduced by Vaidya et al. for learning augmented bloom filters [2], we propose the *Partitioned Learned Count-Min Sketch (PLCMS)*, which is able to take advantage of the full prediction space of a learned model, by partitioning items into different sets, based on multiple predicted frequency thresholds. Each set is handled by a separate Count-Min sketch data structure. We demonstrate that, given fixed partitioning thresholds, the parameters of this PLCMS data structure can be optimized. Empirically, we demonstrate that PLCMS is able to obtain a lower false positive rate for frequent item identification as compared to LCMS and standard Count-Min sketch.

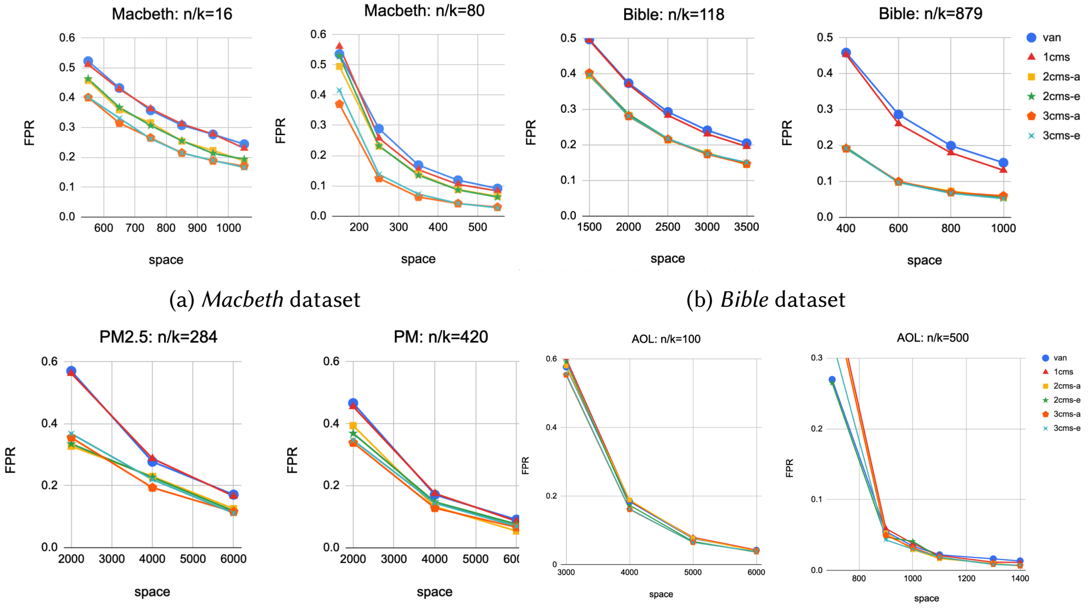
1 INTRODUCTION

Count-Min sketch (CMS) is a space-efficient probabilistic data structure for estimating the frequency of items in a data stream. It is commonly applied to solve the Heavy-Hitters or (ϵ, k)-frequent item problem, in which the goal is to identify all items in a data stream that occur at least n/k times up to a specified error $\epsilon \in (0, 1)$. Here, $k \leq n$ is a parameter and n is the total length of the data stream. In this task, we define the false positive rate (FPR) to be the fraction of infrequent items with frequency less than $(1 - \epsilon)\frac{n}{k}$ that are misidentified as heavy-hitters (i.e., whose estimated frequency in CMS is greater than $\frac{n}{k}$). Recent work has shown that it is possible to improve the CMS frequency estimates by combining a regular CMS with a learned model [1]. The *Learned Count-Min Sketch* (LCMS) uses a learned oracle to predict whether an item is a heavy-hitter. It specifies a single threshold value for the classifier output score space and elements of the data stream that obtain a prediction score above this threshold are placed in an array of *unique buckets*, where each predicted frequent element is assigned its own counter that stores its exact frequency. Items that receive a lower score are placed in a Count-Min sketch data structure, which returns frequency estimates for these items. While LCMS obtains a much lower absolute frequency-estimation error as compared to the standard Count-Min sketch, it does not take the full output score space obtained from the classifier into account.

2 OUR APPROACH

We propose a Partitioned Learned Count-Min Sketch (PLCMS) variant which partitions the items of a data stream according to their frequency scores obtained from the learned model. Items that are predicted to be most frequent have their unique frequencies stored directly and the rest are grouped and assigned to different Count-Min sketch data structures. This allows us to handle items with different frequencies separately and tune the parameters of each group individually to achieve a lower overall FPR. Given a fixed set of thresholds that define regions of the prediction output space, we show how to optimize the space allocations and other parameters for each CMS data structure to minimize the overall false positive rate in the (ϵ, k)-frequent items task. Specifically, we collect a small sample from the stream to estimate the fraction of infrequent elements and the total frequency of items falling in each region. Then, using a convex solver we compute how much space to assign to each region. Finally, we optimize the key parameters of the CMS data structures.

Authors' address: Thuy Trang Nguyen, thuytrangngu@umass.edu; Cameron Musco, cmusco@umass.edu, University of Massachusetts Amherst, Amherst, Massachusetts, USA, 01003.



(a) *Macbeth* dataset

(b) *Bible* dataset

(c) PM2.5 dataset

(d) AOL dataset

Fig. 1. FPR vs. space for various datasets with different $\frac{n}{k}$ values where 'van' denotes the standard CMS and '1cms' is the LCMS variant. The '2cms' and '3cms' labels represent PLCMS with two and three CMS data structures, respectively, where '2cms-a' and '3cms-a' assume the exact number of infrequent elements and total frequency of items in each region, whereas '2cms-e' and '3cms-e' estimate these quantities.

In particular, we set the number of random hash functions used by each CMS according to a derived upper bound formula for the expected false positive rate of PLCMS.

3 RESULTS

We evaluate our algorithm on four datasets: *Macbeth*, *Bible*, Beijing PM2.5 concentration, and AOL search query. Specifically, we use PLCMS with two and three CMS data structures and compare its performance against that of the standard and learned CMS. In Figure 1 we observe that PLCMS outperforms the baseline algorithms on the literary texts and for smaller space sizes on the PM 2.5 dataset. PLCMS does not offer any benefits for AOL since the dataset's distribution is very skewed.

4 FUTURE WORK

The focus of our future work is to tighten the analysis for the upper bound on the expected FPR returned by PLCMS to obtain optimal parameters. In addition, we would like to find a way to optimize the threshold values and the number of partitions with respect to the parameter k.

REFERENCES

[1] Chen-Yu Hsu, Piotr Indyk, Dina Katabi, and Ali Vakilian. 2019. Learning-Based Frequency Estimation Algorithms. In *International Conference on Learning Representations*. https://openreview.net/forum?id=r1lohoCqY7

[2] Kapil Vaidya, Eric Knorr, Tim Kraska, and Michael Mitzenmacher. 2020. Partitioned Learned Bloom Filter. *CoRR* abs/2006.03176 (2020). arXiv:2006.03176 https://arxiv.org/abs/2006.03176

Received 1 May 2023