Technical Perspective: Online Model Management via Temporally Biased Sampling

Ke Yi
HKUST
yike@ust.hk

Randoms sampling from data streams is a problem with a long history of studies, starting from the famous *reservoir sampling* algorithm that is at least 50 years old [2]. The reservoir sampling algorithm maintains a random sample over all data items that have ever been received from the stream. This is not suitable for many of today's applications on evolving data streams, where recent data is more important than older ones.

There are two popular approaches to dealing with evolving data streams in the literature.

Sliding windows: In the sliding window model, only data that have arrived in the window $[now - w, now]$ are relevant, where now is the current time instance, and w is the length (in terms of time) of the window. In the context of random sampling, this means that the sample should be only drawn from data items in the window, each with equal probability. Random sampling in the sliding window model have been well studied, and optimal algorithms are available [1].

Time decay: In the time decay model, the "importance" assigned to each data item decreases as its age. The importance function, in general, can be an arbitrary non-increasing function of age, but the mostly commonly used one is *exponential decay*, where the importance is of an item x is $e^{-\lambda(now-t_x)}$, where t_x is the timestamp of x, and λ is a parameter that controls the rate of the decay. In the context of random sampling, this means that the probability of each item being sampled should be proportional to its importance.

The following figure illustrates the difference between the sliding window model and the time decay model. While the sliding window model applies a sharp threshold (the length of the sliding window) on the age of the items, the time decay model is much "smoother". It gives everyone a chance to be sampled, although older items have exponentially smaller probabilities to get into the sample.

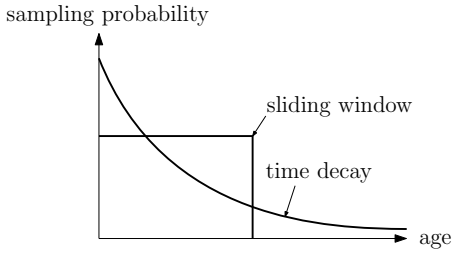

Figure 1: Sampling probability vs. age in the two models.

The following paper by Hentschel, Haas, and Tian takes the time decay approach to random sampling. Building upon prior work, they designed two elegant algorithms with strong theoretical guarantees. The first one, called T-TBS, is very simple to implement and highly scalable, but assumes the arriving batch sizes are i.i.d. with a common mean. The sample size may not be guaranteed when this assumption fails. The second algorithm, called R-TBS, is more complicated, but offers stronger guarantees. It provides a guaranteed upper bound on the sample size, and allows unknown, varying arrival rates. The authors have also applied their time-decayed samples to training machine learning models over evolving data. The results demonstrate promising results showing that these samples can help to refresh the models to capture evolving patterns in the stream more accurately. The paper should be useful for anyone who is interested in random sampling or data analytics over evolving data in general.

1. REFERENCES

[1] V. Braverman, R. Ostrovsky, and C. Zaniolo. Optimal sampling for sliding windows. In *Proc. ACM Symposium on Principles of Database Systems*, 2009.

[2] D. E. Knuth. *Seminumerical Algorithms*, volume 2 of *The Art of Computer Programming*. Addison-Wesley, Reading, MA, 1st edition, 1969.