Scalable Methods for Learning Practical Data-Aware Bloom Filters

Abstract

The approximate set membership problem is a computational problem that involves determining whether a given element belongs to a large set with a certain degree of accuracy or probability. The goal is to minimize the false positive rates while using as few resources (time and space) as possible. This problem arises in various applications, such as data analysis, database management, cryptography, and machine learning. There are several solutions to the set membership problem with false positives, such as Bloom filters, quotient filters, cuckoo filters, and their variants.

Bloom filters have a linear relationship between the size of the filter and its false positive rate guarantees. In recent years, machine learning has become a powerful tool that can significantly improve the performance of Bloom filters. One such improvement is learned Bloom filters (LBF), which use machine learning models to pre-filter the keys and a small backup Bloom filter to handle false negatives. While LBF shows promising performance in some applications, it has some significant drawbacks. In this thesis, we address three of these limitations:

1. LBF demands a fast and small classifier, which is not dependent on negative examples, which can often be hard to come by in applications of LBF. The existing methods fail to meet these criteria.
2. LBF inherits the design problems of machine learning: choice of models and parameters.
3. LBF cannot adapt to changing distributions when keys are inserted into the structure.

To address the first challenge, we developed a random projection-based one-class classifier called Fast Random-projection based One-Class Classifier (FROCC). Our method is based on a simple idea of transforming the training data by projecting it onto a set of random unit vectors that are chosen

uniformly and independently from the unit sphere and bounding the regions based on the separation of the data. A parallel approximation of FROCC, called ParDFROCC, is highly efficient and scales very well with dimensions and the size of data. FROCC achieves up to 3.1 percent points better ROC, with 1.2–67.8× speedup in training and test times over a range of state-of-the-art benchmarks, including the SVM and the deep learning-based models for the OCC task. In addition, we also develop a general framework for analyzing randomized classification algorithms. Using this framework, we prove that FROCC is a stable learning algorithm, that is, it generalizes well with the increase in training data.

We then use the ideas of FROCC to develop a data-aware hash-based Bloom filter called Partitioned Hash Bloom Filter (PHBF) to address the second problem. PHBF works as follows: we partition the Bloom filter into segments, each of which uses a simple projection-based hash function computed using the data. We also provide a theoretical analysis that provides a principled way to select the design parameters of our method: the number of hash functions and the number of bits per partition. We show that it can achieve an improvement in false positive rates of up to two orders of magnitude over standard Bloom filters for the same memory usage, and up to 50% better compression (bytes used per key) for the same FPR, and, consistently beats the existing variants of learned Bloom filters.

Finally, we address the third problem: the adaptability of learned Bloom filters to changing distribution. We propose two distinct approaches for handling data updates encountered in practical uses of LBF: (i) CA-LBF, where we retrain the learned model to accommodate the *unseen* data, resulting in classifier adaptive methods, and (ii) IA-LBF, where we replace the traditional Bloom filter with its adaptive version while keeping the learned model unchanged, leading to an index adaptive method. Our empirical results using a variety of datasets and learned models of varying complexity show that our proposed methods' ability to handle incremental updates is quite robust.