SCALABLE METHODS FOR LEARNING PRACTICAL DATA-AWARE BLOOM FILTERS

ARINDAM BHATTACHARYA

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY DELHI

OCTOBER 2023

SCALABLE METHODS FOR LEARNING PRACTICAL DATA-AWARE BLOOM FILTERS

by

ARINDAM BHATTACHARYA

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Submitted

in fulfilment of the requirements of the degree of Doctor of Philosophy

to the

INDIAN INSTITUTE OF TECHNOLOGY DELHI

OCTOBER 2023

THESIS CERTIFICATE

This is to certify that the thesis titled **Scalable Methods for Learning Practical Data-Aware Bloom Filters**, submitted by **Arindam Bhattacharya**, to the Indian Institute of Technology, Delhi, for the award of the degree of **Doctor of**, is a bona fide record of the research work done by him under our supervision. The contents of this thesis, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

Amitabha Bagchi
Professor
Department of Computer Science and Engg.
Indian Institute of Technology Delhi
New Delhi- 110016

Srikanta Bedathur
Professor
Department of Computer Science and Engg.
Indian Institute of Technology Delhi
New Delhi- 110016

ACKNOWLEDGEMENTS

This thesis is a product of many people's efforts and assistance. First and foremost, I would like to express my gratitude to Amitabha Bagchi and Srikanta Bedathur for supervising me throughout my Ph.D. I was fortunate to have the opportunity to learn from such a unique combination of advisors. From them, I learned how to think like a researcher, express myself clearly, and pay attention to details. From them, I also learned what it means to be a good teacher and an advisor, over and above a Ph.D. supervisor. I am grateful for their support and patience through the uniquely tricky times of the COVID-19 pandemic.

I also want to thank my research committee: Amit Kumar, Syamantak Das, and Sayan Ranu, for reviewing my research progress and giving me constant feedback.

I am fortunate to work with several colleagues throughout my Ph.D. The discussions with them were invaluable to my research. I have benefited tremendously from several long discussions with Sumanth Varambally and Chatur Gudesa.

I am also grateful to the wonderful office staff of the Computer Science and Engineering department. They always have gone out of their way to be helpful; without them, it would have been impossible to maneuver a lot of official affairs as smoothly, especially remotely.

During my stay at IIT Delhi, I have made some excellent friends who made it possible to keep my sanity throughout my Ph.D. The late-night discussions with Vishal Sharma and the lab banter with Sandeep Kumar were perfect antidotes for the stress of graduate school. I am grateful to Dilpreet Kaur and Ovia Seshadri, who often pushed me out of my comfort zone. I also want to thank Ankit Anand and Happy Mittal for their guidance in the initial years.

Finally, I want to thank my parents and sister for their constant and unwavering support throughout my Ph.D. Without them, none of this would be possible.

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.

सार

अनुमानित सेट सदस्यता समस्या एक कम्प्यूटेशनल समस्या है जिसमें यह निर्धारित करना होता है कि क्या कोई दिया गया तत्व एक निश्चित डिग्री की सटीकता या संभावना के साथ एक बड़े सेट में शामिल है या नहीं। लक्ष्य यथासंभव कम संसाधनों (समय और जगह) का उपयोग करते हुए झूठी सकारात्मकता की दरों को कम करना है। यह समस्या विभिन्न अनुप्रयोगों, जैसे डेटा विश्लेषण, डेटाबेस प्रबंधन, क्रिप्टोग्राफी और मशीन लर्निंग में उत्पन्न होती है। झूठी सकारात्मकताओं के साथ सेट सदस्यता समस्या के कई समाधान हैं, जैसे ब्लूम फ़िल्टर, क्वोशैंट फ़िल्टर, कुकू फ़िल्टर और उनके वेरिएंट।

ब्लूम फ़िल्टर में फ़िल्टर के आकार और इसकी झूठी सकारात्मक दर गारंटी के बीच एक लीनियर संबंध होता है। हाल के वर्षों में, मशीन लर्निंग एक शक्तिशाली उपकरण बन गया है जो ब्लूम फिल्टर के प्रदर्शन में काफी सुधार कर सकता है। ऐसा ही एक सुधार सीखा हुआ ब्लूम फ़िल्टर (एलबीएफ) है, जो कुंजी को पूर्व-फ़िल्टर करने के लिए मशीन लर्निंग मॉडल का उपयोग करता है और झूठी नकारात्मकताओं को संभालने के लिए एक छोटे बैकअप ब्लूम फ़िल्टर का उपयोग करता है। यद्यपि एलबीएफ कुछ अनुप्रयोगों में आशाजनक प्रदर्शन दिखाता है, इसमें कुछ महत्वपूर्ण कमियां हैं। इस थीसिस में, हम इनमें से तीन खामियों को सुधरने की कोशिश करते हैं:

1. एलबीएफ एक तेज़ और छोटे क्लासिफायरियर की मांग करता है, जो नकारात्मक उदाहरणों पर निर्भर नहीं है, जिसे अक्सर एलबीएफ के अनुप्रयोगों में प्राप्त करना मुश्किल हो सकता है। मौजूदा तरीके इन मानदंडों को पूरा करने में विफल हैं।
2. एलबीएफ को मशीन लर्निंग की डिजाइन समस्याएं विरासत में मिली हैं: मॉडल और मापदंडों का चयन।
3. जब संरचना में चाबियाँ डाली जाती हैं तो एलबीएफ बदलते वितरण का अनुसरण नहीं कर पाता।

पहली चुनौती का समाधान करने के लिए, हमने एक रैंडम प्रोजेक्शन-आधारित वन-क्लास क्लासिफायरियर विकसित किया, जिसे फास्ट रैंडम-प्रोजेक्शन आधारित वन-क्लास क्लासिफायर (FROCC) कहा जाता है। हमारी पद्धति प्रशिक्षण डेटा को यादृच्छिक इकाई वैक्टर के एक सेट पर प्रक्षेपित करके बदलने के एक सरल आईडिया पर आधारित है जो इकाई क्षेत्र से समान रूप से और स्वतंत्र रूप से चुने जाते हैं और डेटा के पृथक्करण के आधार पर क्षेत्रों को सीमित करते हैं। FROCC का एक समानांतर सन्निकटन, जिसे ParDFROCC कहा जाता है, अत्यधिक कुशल है और आयामों और डेटा के आकार के साथ बहुत अच्छी तरह से मापता है। FROCC ने एसवीएम और ओसीसी कार्य के लिए गहन शिक्षण-आधारित मॉडल सहित अत्याधुनिक बेंचमार्क की एक श्रृंखला में प्रशिक्षण और परीक्षण समय में 1.2-67.8× गति के साथ ROC से 3.1 प्रतिशत अंक बेहतर हासिल किया है। इसके अलावा, हम यादृच्छिक वर्गीकरण एल्गोरिदम का विश्लेषण करने के लिए एक सामान्य रूपरेखा भी विकसित करते हैं। इस ढांचे का उपयोग करते हुए, हम साबित करते हैं कि FROCC एक स्थिर शिक्षण एल्गोरिदम है, अर्थात, यह प्रशिक्षण डेटा में वृद्धि के साथ अच्छी तरह से सामान्यीकरण करता है।

फिर हम दूसरी समस्या के समाधान के लिए पार्टिशन्ड हैश ब्लूम फ़िल्टर (PHBF) नामक डेटा-जागरूक हैश-आधारित ब्लूम फ़िल्टर विकसित करने के लिए FROCC की मूल अवधारणाओं का उपयोग करते हैं। PHBF निम्नानुसार काम करता है: हम ब्लूम फ़िल्टर को खंडों में विभाजित करते हैं, जिनमें से प्रत्येक डेटा का उपयोग करके गणना की गई एक सरल प्रक्षेपण-आधारित हैश फ़ंक्शन का उपयोग करता है। हम एक सैद्धांतिक विश्लेषण भी प्रदान करते हैं जो हमारी पद्धति के डिज़ाइन मापदंडों को चुनने का एक सैद्धांतिक तरीका प्रदान करता है: हैश फ़ंक्शंस की संख्या और प्रति विभाजन बिट्स की संख्या। हम दिखाते हैं कि यह समान मेमोरी उपयोग के लिए मानक ब्लूम फिल्टर की तुलना में परिमाण के दो ऑर्डर तक की झूठी सकारात्मक दरों में सुधार प्राप्त कर सकता है, और समान FPR के लिए 50% तक बेहतर संपीड़न (प्रति कुंजी उपयोग किए गए बाइट्स) और सीखे गए ब्लूम फिल्टर के मौजूदा वैरिएंट्स को हर हाल में पीछे छोड़ता है।

अंत में, हम तीसरी समस्या का समाधान करते हैं: बदलते वितरण के लिए सीखे गए ब्लूम फ़िल्टर की अनुकूलनशीलता। हम एलबीएफ के व्यावहारिक उपयोग में आने वाले डेटा अपडेट को संभालने के लिए दो अलग-अलग दृष्टिकोण प्रस्तावित करते हैं: (i) CA-LBF, जहां हम अनदेखे डेटा को समायोजित करने के लिए सीखे गए मॉडल को फिर से प्रशिक्षित करते हैं, जिसके परिणामस्वरूप क्लासिफायर अनुकूली विधियां होती हैं, और (ii) IA-LBF, जहां हम सीखे गए मॉडल को अपरिवर्तित रखते हुए पारंपरिक ब्लूम फ़िल्टर को उसके अनुकूली संस्करण से बदल देते हैं, जिससे एक सूचकांक अनुकूली विधि बनती है। विभिन्न प्रकार के डेटासेट और अलग-अलग जटिलता से सीखे गए मॉडलों का उपयोग करके हमारे अनुभवजन्य परिणाम बताते हैं कि वृद्धिशील अपडेट को संभालने के लिए हमारे प्रस्तावित तरीकों की क्षमता अच्छी है।

Contents

List of Tables

List of Figures