EFFICIENT AND ACCURATE ALGORITHMS FOR EXTREME CLASSIFICATION

YASHOTEJA PRABHU

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

INSTITUTE OF TECHNOLOGY DELHI

APRIL 2023

EFFICIENT AND ACCURATE ALGORITHMS FOR EXTREME CLASSIFICATION

By

YASHOTEJA PRABHU

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

Submitted

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

to the

INDIAN INSTITUTE OF TECHNOLOGY DELHI

APRIL 2023

Certificate

This is to certify that the thesis titled **Efficient and Accurate Algorithms for Extreme Classification** being submitted by **Mr. Yashoteja Prabhu** for the award of **Doctor of Philosophy** in the Department of Computer Science and Engineering is a record of bona fide work carried out by him under my guidance and supervision at the **Department of Computer Science and Engineering, Indian Institute of Technology Delhi**. The work presented in this thesis has not been submitted elsewhere, either in part or full, for the award of any other degree or diploma unless otherwise stated explicitly. In particular, work in Chapters 3 and 4 were done jointly with other Ph.D. students. Work in Chapter 4 was also done jointly with a Master's student. In each case, the part done by the collaborators appeared in their respective theses.

Manik Varma
Adjunct Professor
Department of Computer Science and Engg.
Indian Institute of Technology Delhi
New Delhi - 110016

Amit Kumar
Professor
Department of Computer Science and Engg.
Indian Institute of Technology Delhi
New Delhi - 110016

Acknowledgements

I am immensely grateful to my advisor Manik Varma who has been an advisor, a mentor, and a confidant to me over the years and to whom I owe most of my professional learnings including the importance of strong work ethics and attention-to-detail. His continued support has been of great help during the difficult phases of my life. I am also thankful to his family with whom I have shared delicious lunches and nice conversations.

I am also deeply grateful to my co-advisor Amit Kumar who was always available for guidance and who has supported me over the years through every step of my Ph.D. This degree could not have been completed without his help.

I am thankful to my research collaborators for contributing to my work and for sharing valuable lessons in coding, experimentation, teamwork *etc.* In particular, I thank Himanshu Jain, Shilpa Gopinath, Anil Kag, Kunal Dahiya, Shrutendra Harsola, Rahul Agrawal, Aditya Kusupati and Nilesh Gupta. I am also grateful to Purushottam Kar who kindly helped me to write my first research paper as well as to review this final thesis. I thank the professors and peers at IIT Delhi for useful technical discussions and suggestions throughout my Ph.D. I thank the CSE office at IIT Delhi for their utmost dedication towards attending to student needs. I especially thank Microsoft for greatly supporting my research work through meaningful collaborations with several product groups. I am indebted to Tata Consultancy Services for funding my Ph.D.; and to MSR India, IARCS and other organizations for providing travel grants to attend conferences and workshops.

This journey would not have been so enjoyable without the constant encouragement and support from my friends at work and in the hostel. I thank Ankit Anand, Arindam Bhattacharya, Anup Bhattacharya, Deepak Ravi, Dinesh Khandelwal, Happy Mittal, Himanshu Jain, Kunal Dahiya, Prachi Jain and many other colleagues who have been my

partners for relishing teas and delicious dinners, exchanging friendly banters and heated arguments, enjoying morning cycle rides and evening frisbee. I find myself extremely fortunate to have had a cohort of such talented and lively peers, and their good nature and dedication to their craft will continue to inspire me. I also thank my co-hostelites, Anil Patidar, Amit Mishra, Arup Roy, Jagatpreet Singh, Samarprit Chakraborty, Soumic Sarkar, and others, for all the tasty Aloo Buns we had together, and the happiness and miseries we jointly lived through. Please forgive me if I have missed anybody. These memories will continue to brighten my days.

I thank the people at MSR India for hosting me during my last few years of Ph.D. and for providing me the much-needed guidance and help. In particular, I thank Kalika Bali for kindly mentoring me over tasty lunches; all the Research Fellows and Research Engineers for collaborating with me on many projects; and the researchers and HR for all their contributions and assistance. I was greatly benefitted by the regularly organized research talks and discussions. I also immensely enjoyed the evening matches of table tennis and foosball and our trips to Sri Lanka and Mahabalipuram.

I thank my chums, Arjun, Manjunath and Rakesh, who never left my side since our school days and provided moral support during my Ph.D. I thank my dear parents and sister for always believing in me and supporting me through all my endeavors. Finally, I thank my lovely and loving wife, Rohini, whose kindness, gregariousness and conviction have left a permanent imprint on me. Her persistent interest in reading my thesis, specifically this acknowledgements section, has been the primary motivation for me to complete this thesis.

Yashoteja Prabhu

Abstract

Classification problems with an extremely large number of fine-grained classes or labels are frequently encountered in present-day applications. For example, data annotation tasks which involve millions of label choices such as Wikipedia categories or Flickr captions are helpful for effective webpage organization and retrieval. Additionally, large-scale recommendation and ranking tasks can also be naturally posed as classification problems by treating each item to be recommended or ranked as a separate label. Extreme classification is a thriving research area of machine learning which studies such large classification problems.

This thesis aims to provide viable algorithmic solutions to extreme classification by addressing its core technical challenges. Extreme classifiers need to train efficiently on millions of data points and labels, and predict in real-time as expected by web-based applications. At the same time, prediction accuracy should be maximized to enhance user satisfaction and revenue. Towards these goals, this thesis develops (1) approaches based on balanced data point partitioning trees whose training and prediction costs scale only logarithmically in the number of labels and hence scale to large datasets, (2) balanced label partitioning tree-based approaches that leverage powerful 1-vs-All classifiers to significantly boost prediction accuracy without compromising efficiency, and finally, (3) an approach based on novel sparse label indices to minimize model size and RAM consumption while ensuring high accuracy and logarithmic efficiency.

This thesis also explores multiple relaxations to the standard extreme classification paradigm in order to effectively leverage the additionally available meta-data. First, it proposes warm-start extreme classification which offers improved predictions for frequent inputs by leveraging label-side features and past interaction logs. Second, it develops the

extreme regression paradigm which relaxes the conventional binary relevance assumption to learn richer models from partial label relevances. Third, it extends extreme classification to handle zero-shot label prediction, a common setting in recommendation tasks.

The algorithms in this thesis are well-suited for real-world applications involving millions of labels. Empirically they were found to outperform existing state-of-the-art extreme classification algorithms on publicly available benchmark datasets. Moreover, in live flights on the Bing search engine, the proposed techniques were also found to significantly improve the key metrics when applied to Sponsored and Dynamic Search Ad applications

सार

आज के अनुप्रयोगों में बहुत बड़ी संख्या में सुक्ष्म वर्गों या लेबलों के साथ वर्गीकरण की समस्याएं अक्सर सामने आती हैं। उदाहरण के लिए, डेटा टिप्पणी कार्य जिसमें विकिपीडिया श्रेणियों या फ़्लिकर कैप्शन जैसे लाखों लेबल विकल्प शामिल होते हैं जो प्रभावी डेटा संगठन और पहुँच के लिए सहायक होते हैं। इसके अलावा, बड़े पैमाने पर चर संज्ञा और श्रेणी कार्यों को भी स्वाभाविक रूप से वर्गीकरण समस्याओं के रूप में पेश किया जा सकता है, प्रत्येक वस्तु को एक अलग लेबल के रूप में अनुशंसित या श्रेणी कर सकते हैं। चरम वर्गीकरण मशीन लर्निंग का एक उभरता हुआ शोध क्षेत्र है जो इस तरह की बड़ी वर्गीकरण समस्याओं का अध्ययन करता है।

इस निबंध का उद्देश्य चरम वर्गीकरण के मूल तकनीकी चुनौतियों के हल के लिए व्यवहार्य एल्गोरिथम समाधान प्रदान करना है। चरम वर्गीकारक को लाखों डेटा बिंदुओं और लेबलों पर कार्य करने के लिए कुशलता से प्रशिक्षित करने की और वेब-आधारित अनुप्रयोगों द्वारा अपेक्षित समय में निष्कर्ष की आवश्यकता होती है। इसके अतिरिक्त, उपयोगकर्ता संतुष्टि और राजस्व को अधिकतम करने के लिए निष्कर्ष यथार्थता अधिक होनी चाहिए। इन लक्ष्यों की ओर, यह निबंध निम्नलिखित विकसित करता है: (1) संतुलित डेटा बिंदु विभाजन वाले पदानुक्रम पर आधारित दृष्टिकोण जिनके प्रशिक्षण और निष्कर्ष की लागत केवल लेबल की संख्या में लॉगरिदमिक रूप से बढ़ती है और इसलिए बड़े डेटासेट के पैमाने पर चलती है, (2) संतुलित लेबल पदानुक्रम-आधारित दृष्टिकोण जो मजबूत 1-विरूद्ध-सभी वर्गीकारक की लाभ उठाते हैं और, यथार्थ और मितव्ययी निष्कर्ष करते हैं, (3) उच्च यथार्थता और लॉगरिदमिक मितव्य सुनिश्चित करते हुए, मॉडल परिमाण और रैम की खपत को कम करने के लिए नई और विरल लेबल सूचकांकों पर आधारित एक दृष्टिकोण।

यह निबंध उपलब्ध मेटा-डेटा का प्रभावी ढंग से लाभ उठाने के लिए चरम वर्गीकरण में उपयोग की जाने वाली समस्या में विभिन्न छूटों की भी पड़ताल करती है। सबसे पहले, यह वार्म-स्टार्ट चरम वर्गीकरण का प्रस्ताव करता है जो बेहतर निष्कर्ष के लिए प्रकट लेबल डेटा का लाभ उठाता है। दूसरा, यह चरम रिग्रेशन विकसित करता है जो आंशिक लेबल प्रासंगिकता से समृद्ध मॉडल सीखने के लिए पारंपरिक बाइनरी प्रासंगिकता धारणा को शिथिल करता है। तीसरा, यह शून्य-शॉट लेबल को संभालने के लिए चरम वर्गीकरण का विस्तार करता है।

इस निबंध में प्रस्तावित चरम वर्गीकरण तकनीक लाखों लेबल वाले वास्तविक दुनिया के अनुप्रयोगों के लिए उपयुक्त हैं। प्रयोगों में, वे सार्वजनिक रूप से उपलब्ध बेंचमार्क डेटासेट पर मौजूदा अत्याधुनिक चरम वर्गीकरण एल्गोरिदम को बेहतर प्रदर्शन किया। इसके अलावा, बिंग सर्च इंजन पर लाइव फ्लाइट्स में, प्रस्तावित तकनीकों को प्रायोजित और

डायनामिक खोज विज्ञापन अनुप्रयोगों पर लागू करने पर प्रमुख मैट्रिक्स में महत्वपूर्ण सुधार करने के लिए भी पाया गया।

Contents

List of Figures

List of Tables