EXTREME CLASSIFICATION: LOSS FUNCTIONS, ALGORITHMS AND APPLICATIONS

HIMANSHU JAIN

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY DELHI
MARCH 2020

EXTREME CLASSIFICATION: LOSS FUNCTIONS, ALGORITHMS AND APPLICATIONS

by

HIMANSHU JAIN

Department of Computer Science and Engineering

Submitted

in fulfillment of the requirements of the degree of Doctor of Philosophy

to the

Indian Institute of Technology Delhi

MARCH 2020

Certificate

This is to certify that the thesis titled **EXTREME CLASSIFICATION: LOSS FUNCTIONS, ALGORITHMS AND APPLICATIONS** being submitted by **MR. HIMANSHU JAIN** for the award of **Doctor of Philosophy** in **Computer Science and Engineering** is a record of bona fide work carried out by him under our 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, the work done in Chapter 2 was done jointly with a PhD student. Work done in Chapters 3, 4 and 5 was done jointly with research fellows at Microsoft Research. In each case, the part done by the collaborators appeared in their respective theses.

Rahul Garg
Professor
Department of Computer Science & Engg.
Indian Institute of Technology Delhi
New Delhi- 110016

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

Acknowledgements

I am deeply thankful to my advisor Dr Manik Varma for his guidance and support throughout this thesis. His passion towards the subject, drive to do something impactful and ability to pay attention to the minutest of details always kept me motivated, on toes and pushed me to do my best. I have also vastly benefited from his unique mix of industry and academia experience. I would certainly not have hoped for a better learning experience than this. I would also like to acknowledge the help and support that I received from Prof. Rahul Garg.

I am also thankful to Prateek Jain and Purushottam Kar for their invaluable inputs throughout my research. I am very grateful to Nicolas Mayoraz for hosting me for the summer internship at Google. It was an amazing learning experience and offcourse I can never forget the half-dome hike. I am also thankful to my collaborators Kush Bhatia, Venkatesh Balasubramanian, Bhanu Chunduri, and Yashoteja Prabhu. I got to learn a lot from each one of you.

I am extremely thankful to Google for their generous financial support throughout my PhD. Because of their support, not only I was able to financially sustain throughout these years, but I could also easily attend conferences and workshops anywhere in the world without worrying about arranging funds. I would also like to acknowledge the administrative staff at the computer science department in IIT Delhi for their continuous help throughout these years.

It is almost impossible to think of successfully completing the PhD without the social support of friends. I was lucky enough to have an amazing group of friends at IIT Delhi. I am particularly grateful to Ankit Anand, Dinesh Khandelwal, Prachi Jain, Happy Mittal

and Yashoteja Prabhu for the endless laughs, heated political discussions, some amazing trips and numerous eat outs (thanks to the lovely mess food). I am also thankful to Kunal, Saurabh, Ashish, Dilpreet, Anup and many other fellow PhD students who made the stay at IIT Delhi a memorable one. I am also thankful to my friends outside IIT Delhi, particularly Varun, Aditya, Anuj, Brajesh and Ayush as they kept visiting and always paid for the restaurant bills. One thing that I eagerly looked forward to every year and that always replenished me was treks. So I am thankful to Vivek, Rahul, Shashank, Aditya and Ankit for the incredible trek experiences.

Finally, I can't thank my parents enough for the numerous sacrifices that they made so that I can get a good eduction. Without their support, encouragement and unflinching faith in me, none of this would have been possible. I am also thankful to my sister and brother-in-law for always supporting me and to all my family members who never asked me the dreaded question – *When are you finishing your PhD?*

Himanshu Jain

Abstract

The objective in extreme multi-label learning is to learn a classifier that can automatically tag a data point with the most relevant subset of labels from an extremely large label set. Extreme multi-label classification is an important research problem as it not only lets us tackle web-scale classification problems but it has also opened a new paradigm for solving ranking and recommendation problems. This thesis focusses on developing scalable algorithms and appropriate loss functions for extreme classification problems which can lead to state-of-the-art performance on large-scale ranking and recommendation applications. In particular, it makes the following contributions –

1) It proposes loss functions suitable for extreme multi-label learning that do not erroneously treat missing labels as irrelevant but instead provide unbiased estimates of the true loss function even when ground truth labels go missing under arbitrary probabilistic label noise models. These loss functions also naturally promote the accurate prediction of infrequently occurring, hard to predict, but rewarding tail labels.

2) It develops the SLEEC algorithm which is an embedding based extreme multi-label learning algorithm. SLEEC addresses some of the major limitations of embedding based methods such as high training and prediction cost and low prediction accuracy.

3) It develops the Slice algorithm for extreme multi-label learning with low-dimensional dense features that scales to 100 million labels and 240 million training points.

4) It reformulates the problem of recommending related queries on a search engine as an extreme classification task and demonstrates that Slice could significantly improve recommending related searches on Bing.

सार

चरम मल्टी-लेबल सीखने का उद्देश्य एक क्लासिफायरियर सीखना है जो स्वचालित रूप से एक अत्यंत बड़े लेबल सेट से सबसे प्रासंगिक सबसेट के साथ डेटा बिंदु को टैग कर सकता है। एक्सट्रीम मल्टी-लेबल वर्गीकरण एक महत्वपूर्ण शोध समस्या है क्योंकि यह न केवल हमें वेब-स्केल वर्गीकरण समस्याओं से निपटने में मदद करता है, बल्कि इसने रैंकिंग और अनुशंसा समस्याओं के समाधान के लिए एक नया प्रतिमान भी खोला है। यह थीसिस स्केलेबल एल्गोरिदम को विकसित करने और चरम वर्गीकरण समस्याओं के लिए उचित नुकसान के कार्यों पर ध्यान केंद्रित करती है, जिससे बड़े पैमाने पर रैंकिंग और सिफारिश अनुप्रयोगों पर अत्याधुनिक प्रदर्शन हो सकता है।
विशेष रूप से, यह निम्नलिखित योगदान देता है -

1) यह चरम मल्टी-लेबल सीखने के लिए उपयुक्त नुकसान कार्यों का प्रस्ताव करता है जो लापता लेबल को अप्रासंगिक नहीं मानते हैं, लेकिन इसके बजाय जमीनी सच्चाई लेबल के मनमाने ढंग से लेबल शोर मॉडल के तहत गायब होने पर भी वास्तविक नुकसान फ़ंक्शन के निष्पक्ष अनुमान प्रदान करते हैं। ये नुकसान कार्य स्वाभाविक रूप से होने वाली सटीक भविष्यवाणी को बढ़ावा देते हैं, भविष्यवाणी करना कठिन है, लेकिन पूंछ को पुरस्कृत करते हैं।

2) यह SLEEC एल्गोरिथ्म को विकसित करता है जो एक एम्बेडिंग आधारित मल्टी-लेबल लर्निंग एल्गोरिथम है। SLEEC उच्चतर प्रशिक्षण और भविष्यवाणी लागत और कम भविष्यवाणी जैसे अंतर्निहित तरीकों की कुछ प्रमुख सीमाओं को संबोधित करता है।

3) यह कम आयामी सघन सुविधाओं के साथ चरम मल्टी-लेबल सीखने के लिए स्लाइस एल्गोरिदम विकसित करता है जो 100 मिलियन लेबल और 240 मिलियन प्रशिक्षण बिंदुओं को मापता है।

4) यह खोज इंजन पर संबंधित प्रश्नों की सिफारिश करने की समस्या को एक चरम वर्गीकरण कार्य के रूप में सुधारता है और यह दर्शाता है कि बिंग पर संबंधित खोजों की सिफारिश करने में स्लाइस काफी सुधार कर सकता है।

Contents

List of Figures

List of Tables