EFFECTIVE INFERENCE & LEARNING IN STATISTICAL RELATIONAL MODELS

HAPPY MITTAL

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY DELHI

NOVEMBER 2021

EFFECTIVE INFERENCE & LEARNING IN STATISTICAL RELATIONAL MODELS

by

HAPPY MITTAL

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

NOVEMBER 2021

Certificate

This is to certify that the thesis titled **Effective Inference & Learning in Statistical Relational Models** being submitted by **Mr. Happy Mittal** for the award of **Doctor of Philosophy** in **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, 5, and 6 was done jointly with undergraduate students. Work in some parts of Chapters 4 and 5 was also done jointly with master's students. In each case, the part done by the collaborators appeared in their respective bachelor's or master's theses.

Dr. Parag Singla
Associate Professor
Department of Computer Science and Engineering
Indian Institute of Technology Delhi
New Delhi- 110016

Acknowledgements

I wish to thank my advisor Dr. Parag Singla, who guided me in this journey of Ph.D. His command over the subject helped me get started very early in my Ph.D. career. He taught me how to think clearly about the problem, and how to write down ideas floating in mind to bring out a research paper. I remember him saying in my first year that I will earn my Ph.D. only when I will be able to write a full research paper by myself. I think this is a very good benchmark for anyone who wants to analyze himself/herself whether he/she is ready to finish a Ph.D. Moreover, Parag is a great human being. I have learned a lot of traits from him and I wish to follow his footsteps rest of my life also. I would also like to thank Dr. Mausam, who at various times, gave honest and critical feedback on my projects. He also helped me decide the career path after the Ph.D. and presented both pros and cons of the options I had. I am also grateful to Dr. Vibhav Gogate, with whom we collaborated for all our research projects, and that collaboration turned out to be great. He has a great knowledge of his subject, which was reflected in the various meetings I have had with him. I also thank Somdeb Sarkhel with whom I exchanged many research ideas over the emails, and he always helped me with his vast knowledge in the area of SRL.

I am grateful to my parents and my brother Nicky Mittal who supported me throughout this journey. I also want to thank my wife Priyanka who constantly supported me during my endeavour. Moreover, the kids (Akshita and Samaira) in my family bugged me a lot during my thesis writing. Finally, a special mention to my little one Aradhya because of whom I had to spend many sleepless nights.

During my PhD life, I developed a great friendship with my Ph.D. colleagues: Ankit Anand, Prachi Jain, Vishal Sharma, Yashoteja Prabhu, Himanshu Jain, Dilpreet Kaur, Dinesh Khandelwal, and many more with whom I had both technical and non-technical discussions, helping me reduce the stress of academic life.

I worked with many bachelor's and master's students during my research, and because of their support and hard work, I was able to publish good quality work. I would like to thank Prasoon Goyal, Anuj Mahajan, Shubhankar Singh, Noman Sheikh, Piyush Gupta, and Ayush Bhardwaj who spent numerous sleepless nights working and experimenting in labs.

I would also like to thank Prof. Kristian Kersting who hosted me for a 3 months research internship at Dortmund University. I learned a lot from him, and also got a flavor of how research is conducted outside Indian universities.

Finally, I am thankful to Tata Consultancy Services for funding my Ph.D. during the past 4 years that helped me attending and presenting my work at various international conferences.

Happy Mittal

Abstract

A vast majority of traditional Machine Learning algorithms make the i.i.d. assumption, i.e., instances are assumed to be sampled independently from an identical distribution. However, the real world data is relational i.e. there exist various types of relations among the instances which need to be captured. Moreover, these relations also have a certain degree of uncertainty that needs to be modeled. The field of Statistical Relational Learning (SRL) achieves the merger between the two by explicitly modeling relations as well as uncertainty among them. Due to the underlying template structure, SRL models often result in the construction of large ground networks posing serious computational challenges. Therefore, designing effective inference and learning algorithms becomes critical for the success of these models. This thesis focuses on proposing multiple such inference and learning techniques: (a) We first propose two different algorithms for lifted inference, i.e., techniques which exploit the symmetry of the underlying relational representation. In the first approach, we exploit the underlying symmetry using a rule called Single Occurrence (SO) for MAP inference. SO rule results in inference time which is independent of domain size under certain conditions. Second, we propose a language for specifying constraints

for constructing the underlying ground theory and extend existing lifting techniques to work with these constraints.

(b) In the second part of this thesis, we present two different techniques for effective learning in SRL models. First, we present an algorithm for fine-grained learning, where we learn a different weight for different subsets of the groundings of the same logical formula. We show that this can result in improved prediction accuracy by reducing bias. Second, we present a novel re-parameterization technique to generalize the weights learned for logical formulas across domains of varying sizes. We also present various theoretical properties of our formalism.

In each of the above cases, we present experimental results validating the efficacy of proposed algorithms on a mix of real and artificial datasets. Much of our work is presented in the context of Markov Logic Networks (MLNs), which is a popular SRL formalism though our techniques are general enough and can be extended to other similar representations. We conclude the thesis with directions for future work.

सार

पारंपरिक मशीन लर्निंग एल्गोरिथ्म का एक विशाल बहुमत i.i.d. धारणा, यानी, उदाहरणों को एक समान एवं स्वतंत्र रूप से वितरित मानता है । यद्यपि वास्तविक विश्व डेटा संबंधपरक है अर्थात उदाहरणों के बीच विभिन्न प्रकार के संबंध मौजूद हैं जिन्हें पकड़ने की आवश्यकता है। इसके अतिरिक्त, इन संबंधों में कुछ कुछ मात्रा में अनिश्चितता भी होती है जिसे प्रतिरूपित करने की आवश्यकता होती है। स्टैटिस्टिकल रिलेशनल लर्निंग (एसआरएल) का क्षेत्र स्पष्ट रूप से मॉडलिंग संबंधों के साथ-साथ उनके बीच अनिश्चितता के बीच विलय को प्राप्त करता है। अंतर्निहित टेम्पलेट संरचना के कारण, एसआरएल मॉडल अक्सर बड़े ग्राउंड नेटवर्क का निर्माण करते हैं जो गंभीर कम्प्यूटेशनल चुनौतियों का सामना करते हैं। इसलिए, इन मॉडलों की सफलता के लिए प्रभावी इन्फेरेंस और लर्निंग के एल्गोरिथ्म को डिजाइन करना महत्वपूर्ण हो जाता है। यह थीसिस कई ऐसे इन्फेरेंस और लर्निंग की तकनीकों को प्रस्तावित करने पर केंद्रित है: (ए) हम पहले लिफ्टेड इन्फेरेंस के लिए दो अलग-अलग एल्गोरिथ्म का प्रस्ताव करते हैं, यानी ऐसी तकनीकें जो अंतर्निहित संबंधपरक निरूपण की समरूपता का फायदा उठाती हैं। पहले दृष्टिकोण में, हम एमएपी अनुमान के लिए सिंगल अकर्रेंस (एसओ) नामक नियम का उपयोग करके अंतर्निहित समरूपता का फायदा उठाते हैं। SO नियम का परिणाम इन्फेरेंस समय होता है जो कुछ शर्तों के तहत डोमेन आकार से स्वतंत्र होता है। दूसरा, हम अंतर्निहित ग्राउंड थ्योरी के निर्माण के लिए बाधाओं को निर्दिष्ट करने के लिए एक भाषा का प्रस्ताव करते हैं और इन बाधाओं के साथ काम करने के लिए मौजूदा लिफ्टिंग की तकनीकों का विस्तार करते हैं। (बी) इस थीसिस के दूसरे भाग में, हम एसआरएल मॉडल में प्रभावी लर्निंग के लिए दो अलग-अलग तकनीकों को प्रस्तुत करते हैं। सबसे पहले, हम अच्छी तरह से लर्निंग के लिए एक एल्गोरिथ्म प्रस्तुत करते हैं, जहां हम एक ही तार्किक सूत्र के आधार के विभिन्न उपसमुच्चय के लिए एक अलग वजन सीखते हैं। हम दिखाते हैं कि इससे पूर्वाग्रह को कम करके भविष्यवाणी की सटीकता में सुधार हो सकता है। दूसरा, हम अलग-अलग आकार के डोमेन में तार्किक सूत्रों के लिए सीखे गए भार को सामान्य बनाने के लिए एक नया पुन: पैरामीटरीकरण तकनीक प्रस्तुत करते हैं। हम अपने सूत्रवाद के विभिन्न सैद्धांतिक गुण भी प्रस्तुत करते हैं। उपरोक्त प्रत्येक मामले में, हम वास्तविक और साथ ही कृत्रिम डेटासेट पर प्रस्तावित एल्गोरिथ्म की प्रभावकारिता को मान्य करते हुए प्रयोगात्मक परिणाम प्रस्तुत करते हैं। हमारा अधिकांश काम मार्कोव लॉजिक नेटवर्क्स (एमएलएन) के संदर्भ में प्रस्तुत किया गया है, जो एक लोकप्रिय एसआरएल सूत्रवाद है, हालांकि हमारी तकनीक काफी सामान्य है और इसे अन्य समान निरूपणों तक बढ़ाया जा सकता है।

Contents

List of Figures

List of Tables