Possible topics for student work in statistical relational AI

Statistical Relational Artificial Intelligence (StaRAI) is an exciting field within contemporary Artificial Intelligence. By combining the highly sophisticated reasoning capabilities of Logical AI with the flexibility of modern Statistical AI in a single organic framework, StaRAI can be applied to highly structured and yet uncertain data, which is a challenge for conventional frameworks of either tradition. Existing applications include bioinformatics and Natural Language Processing, and due to its transparent presentation via (probabilistic) logical rules, it fits very well into the framework of explainable artificial intelligence. The potential of StaRAI in the form of Markov logic networks, one of its main formalisms, is explained for a general audience in this piece from the online journal datanami: www.datanami.com/2018/07/03/can-markov-logic-take-machine-learning-to-the-next-level

Those topics that I consider would also be suitable for Bachelor students are marked with a star (*)

General Practical StaRAI

1. Using StaRAI to model the local epidemiology of infectious diseases: a very topical issue for universities these days is the impact that lecture attendance has on the spread of infectious diseases, and how changes to the way university teaching is administered can change that. In order to give a reasonable answer to that question, one has to first model student attendance in the first place. Then one has to add some layer of epidemiological information, such as “how much more likely is one to get an illness if one shares a lecture with someone who has it”. Finally, one has to model how student attendance and disease transmission would change under imposing measures, such as “no lectures of more than 100 people”, “lectures of more than 100 people only online” etc. Due to the complex, counterfactual reasoning required here, StaRAI is perfectly suited for this challenge. *

2. Using StaRAI for bridge bidding: learning patterns for hand evaluation and then making an opening bid: The card game “bridge” has proven very challenging for artificial intelligence, mainly since it is a game of imperfect information and partnership cooperation. Furthermore, there is a requirement in the laws of bridge to disclose any agreements that you have to your opponents, explicitly making a demand of explainability on a computer pair. This makes it ideally suited for statistical relational AI, since this framework can combine probabilistic reasoning with clear explainability. A perfect starting problem would be to decide on an opening bid. *

Scaling of statistical relational AI (Practical)

My recent work has focused on understanding how Statistical Relational approaches scale with changing domain sizes. This is very important, for instance, when faced with random samples on which to train a model, or in applications that naturally come with different domain sizes (e. g. schools with different numbers of classes and pupils). It has become clear that in many, though not all, cases it will be appropriate to scale the parameters of a model with the domain sizes. In my work I have derived a formalism (DA-RLR) which successfully implements this idea for the underlying formalism of Relational Logistic Regression (RLR), which is based on (directed) Bayesian networks, while previous work has focused on Markov Logic Networks (MLN) and its scaled analogue DA-MLN.


As scaling is a major issue in a world faced with the challenges of big and complex data, there are many interesting possibilities associated with this work:

1. Applying DA-RLR to the IMDB and the WebKB and compare performance to that reported for DA-MLN: On those commonly used model datasets, DA-MLN have been found to significantly outperform unscaled MLN when faced with random training samples. It would be very interesting to implement a scaled DA-RLR and an unscaled RLR approach on these datasets and then compare them both to each other and to the MLN and DA-MLN results reported in the literature. *
2. Finding real life examples of scenarios in which scaled or unscaled models are expected to be preferable and evaluating the effectiveness of RLR and DA-RLR on those (several projects feasible): By thinking of scaled weights as proportions, one can predict from the nature of the application whether scaling by domain size would be appropriate or not. However, it would be very interesting to actually evaluate scaled, unscaled and partially scaled approaches on datasets where they are predicted to perform best. This would be a very valuable contribution to the problem of choosing the correct representation formalism for a real-life application! *

3. In particular, take large datasets from the StaRAI literature and train the formalism on small random samples: The most promising application of DA-RLR is training StaRAI on a random sample rather than on an entire dataset. Due to the high algorithmic complexity of inference and learning in StaRAI, random sampling could be crucial for making large datasets accessible to statistical relational learning. Therefore, it would be very interesting to take large datasets from the literature and compare the performance of a DA-RLR that is trained on a random sample to a StaRAI model that is trained on the entire dataset. *

4. Evaluating the performance of RLR as opposed to MLN: Since our scaling model relies on RLR as the underlying formalism, it would be great to evaluate how well RLR actually performs when compared to MLN, assuming that state-of-the-art learning algorithms are used for both representations. While there is some literature on this, it would be very interesting to see whether there is some correspondence between features of the data set used and the preferable representation.

Scaling of statistical relational AI (Theoretical)

1. Asymptotic Probabilities in first-order logic and infinite models of MLN or RBN: In finite model theory, much of the literature on 0-1 laws and asymptotic probabilities in first-order theories is based on infinite models of several suitable extensions of the theory. It seems very plausible that one could reframe many of those results using infinite models of a single theory of MLN (or a directed analogue) instead. This would be a beautiful bridge back from cutting-edge StaRAI to classical finite model theory.

2. Prove that DA-MLN are well-behaved when they are like DA-RLR (homogeneous formulas): The representation of asymptotic probabilities in DA-RLR rests on there being just a single dependent formula to consider in every condition, and that therefore no aggregation function is required. However, if every formula involved in an undirected model has the same number of connections, the aggregation functor would again be trivial and therefore one should be able to derive the same conclusion. This would be interesting because it would be the first general result on asymptotic behaviour in DA-MLN. *

3. Scaling in ProbLog: Problog, a StaRAI formalism based on logic programming, is one of the most developed directed approaches to StaRAI, and has advantages i. e. when dealing with cycles in the directed model. However, since it is not parametrised by weights, one cannot just copy and paste the “DA-” concept onto ProbLog. It would therefore be particularly interesting to undertake some study on the scaling behaviour in this approach and perspective come up with some analogue of weight-scaling for this formalism.

4. Coloured DA-RLR: An interesting way to avoid cycles in the grounding of a relational belief network is simply to forbid them. This can be implemented for instance by colour-coding the nodes. However, when moving to asymptotic probabilities in DA-RLR, we obtain a closed representation that is no longer obviously related to a fixed grounding of the variable domains. It would thus be very interesting to see how one could carry over the idea of colour coding the nodes to control groundings to the limit representation.

5. Multi-valued RLR: In practice, many domains are multi-valued rather than binary, and therefore having a multi-valued version of relational logistic regression (and perspective, of DA-RLR) would be very valuable. This has been done successfully in the non-relational case, but is yet to be done for RLR.