Topic 3: Caveat Lector

Assignment 3
Topics in Algorithmic Data Analysis SS'17
Submitted to: Dr. Jilles Vreeken
Due by: June 29, 2017 before 10.00 hours

Name: Harshita Jhavar
Matriculation Number: 2566267
eMail: s8hajhav@stud.uni-saarland.de

Introduction

Causal inference from observational data is a very fundamental problem in data mining. Given data over a joint distribution, the effects between a pair of random variables X and Y could be one such interesting settings to identify the causal direction X→Y or Y→X. Many different approaches have been proposed in this direction. One of them is discussed in Vreeken[1] in which the author describes a causal inference rule based on Kolmogorov complexity. The implementation of this rule is presented as ERGO, which is based on cumulative and Shannon entropy. In this report, I will highlight the different choices made by the author during the course of the design of this algorithm and their repercussions. Further, I will pen down my suggestions for improvements on the same ground. Later, I will discuss how the mining algorithm by Heikinheimo[2] for the low entropy sets and trees represent causality. In the end, I will conclude with a summarisation of all the remarks made in this report.

An algorithmic design is about choices and intentions, it is not accidental!

In Vreeken[1], I found the key idea as an intuitive approach to determine the causal direction between a pair of random variables. If Y is caused by X, describing Y using X is easier. If X causes Y, the shortest description of the joint distribution P(X,Y) is given by the separate descriptions of P(X) and P(Y|X) and these two descriptions will be less dependent than P(Y) and P(X|Y). However, there are some computable approximations made which involve arbitrary choices as Kolmogorov complexity is not computable. Some of these choices are:

- The author prefers to approximate the Kolmogorov complexity using entropy. There should be lower entropy in the causal direction or in the right direction as compared to the wrong direction. There could be other measures also for the same, for example, mutual information, MDL. Unlike mutual information, entropy is in an asymmetric form, which makes it possible to measure cause-and-effect relationships. Thus, author prefers entropy over mutual information.

- The design of this rule is based on the assumption of causal sufficiency, i.e., there is no hidden object Z which might be causing X and Y. I agree that while this choice of assumption narrows down the problem to determining if X → Y or Y → X, still, this does not generalize the rule to a real world scenario in which it is totally possible that one agent could participate as a cause for more than one object or vice versa. Why should we be restricted with dealing only univariate pairs when in real complex world, there are lot many multivariate dependencies.

- I find that ERGO instantiates its framework with cumulative entropy to infer the causal direction but it is only restricted to the continuous real valued data. Causal inference from discrete data has not received attention here.

- To cope with objects of different complexities and to avoid the bias of the inference of the causal direction based on the absolute difference of K(Y|X') and K(X|Y'), the choice of normalisation has been made in the algorithm. Well, it now considers the difference in relative conditional complexity to find the causal direction of information. I find a very strong assumption made here. What if the observations recorded for X or Y are dynamically increasing i.e. are not yet complete and contains only a small number of values while the other values are collected incrementally as soon as they arrive in the system. In this case the upper and lower bounds of the values cannot be estimated close to the real values.

- I believe that the choice of considering $\Delta_{X' \to Y}$ instead of $\Delta_{X \to Y}$ explicitly for better or worse, steers the process to the subjective goals. While considering $\Delta_{X' \to Y}$, we measure the information provided by the model X' and Y' which is the best generalisation for X and Y. However, I am still dubious on how generalised this version of representation is if we are ignoring noise or the randomness that may be present in the observed data. One cannot just ignore randomness in the data, instead, one needs to find out a way to deal with it. For example by using Additive Noise Models as done in Peters[4]. Otherwise, we are making the complex task of representation of the information of the data easier by missing the fun part of trading-off with the noise in it. Inference ignoring the uncertainty due to the covariate selection may be overoptimistic. While $\Delta_{X' \to Y}$ is what suits an ideal scenario, $\Delta_{X \to Y}$ is what represents the real scenario.

- To estimate the conditional cumulative entropy for X and Y, the design choice is to discretise the dimension only in the previous step and select dimension X_i with the minimum entropy. I agree that this increases the efficiency of the algorithm and reduces the measurement of the model complexity. However, this again is another choice in order to make the road towards the subjective goal smoother. There is no sufficient proof which confirms that the dimension of the minimum entropy is enough as a choice for modeling. Intuition said so is never a valid answer. Moreover, if the Markov property i.e a certain length of the past on which there is dependence, is not satisfied, entropy may fail to measure the causal relations in the system.

List of improvements!

Based on the different choices made during the rule-design for causal inferencing using Kolmogorov complexity and some literature survey, here I discuss some of the improvements on the same.

- In Vreeken[1], the discrete data i.e the categorical or binary was not considered. Causal inference from discrete data could be handled by using Minimum Description Length(MDL). In Budhathoki et.al.[3] and Budhathoki et.al.[5], the authors describe their experiments in this line of research based on the algorithmic Markov condition using stochastic complexity and conclude it as better performing than most other state of the art methods. Moreover, now we are not limited only to real valued continuous data only.

- In Vreeken[1], ERGO is limited to univariate pairs only. This limitation can again be improved by using MDL (inspired from Bertens et.al.[8]) as the basis of the approximation to come up with a causal model as done in Alexander et.al.[6]. Dealing with multivariate random variables is moving us closer to the real world scenario. Another challenge in this line of research is that we have to assume that all confounding variables are observed.

However, I believe that the choice of covariates to control for is primarily based on subject matter knowledge. It may result in a large covariate vector in the attempt to ensure that all confounding variables are present. However, including redundant covariates can affect bias and efficiency of nonparametric causal effect estimators, e.g., due to the curse of dimensionality. In Persson et.al.[7], the authors have come up with a very interesting framework which is based on dimension reduction to perform a backward elimination procedure assessing the significance of each covariate. Thus, there could be improvement in this line of research.

- It appears to me that one can also initiate in the direction of causal structure learning from the data. This will not only extend this framework to the real or discrete valued data but can also be extended to textual data. For example, in a knowledge graph representation of a characters of a movie, one could find the direction of causality based on the causal structural learned model. This latter part interests me the most. One could also do sentiment analysis on these fictional characters to determine the causality of the conspiracy in a story.

- One can also think of extending the framework to time series data so that temporal dependencies could be noted. This is again applicable for coming up with a timeline representation of the evolved sentiments in a fiction. This is also applicable for any kind of analogous signal analysis.

- To be sure of whether the given data points under observation show any dependence or not, capturing correlation in subspaces is another measure which could be used to get the right set of variables to focus on in the given set of observations. I completely understand that while correlation doesn't imply causality but causality imply correlation (not always linear) as for A to be a cause of B, they must be associated in some way. Thus, we can think of this direction to find a measure of capturing correlation to reduce the dimension of the input and get the right set of variables to focus on.

Bonus - Do low entropy sets and trees exhibit some form of causality representation?

From the discussion above, I am clear that there should be low entropy in the right direction of cause as compared to the wrong direction. Thus, there can be line of research to work on developing algorithms which claim to generate through data mining such data structures which identify low entropy sets and trees from the raw form of input data. In Heikinheimo et.al.[2], the authors describe two such algorithms implemented on binary data - U-tree mining algorithm (edges directed towards the root) and D-tree mining algorithm (edges directed away from the root). The idea is that this definitely exhibits some form of causality representation because these directed trees identify all low entropy combinations in the dataset. I believe, this results in reduction of the dimension of the input data and gives us the set of variables to focus on to determine different possible causality associations. Since, we have combinations with lower entropy, we can definitely conclude that if the direction in two different nodes of a low entropy tree, say A$\rightarrow$B where A is at higher level than B, then, there is a higher possibility that B is caused by A. This is a good way to filter the set of variables to focus on from a large universe of possible input variables. I believe, while the U-tree and D-tree are currently restricted to binary datasets only, it will be much more interesting to come up with a hybrid approach. One line of research could be to determine the low entropy sets from input data using U-tree or D-tree algorithm and then, on this reduced sets of low entropy data, one could use approximations for

Kolmogorov complexity (by MDL or by Shannon/cumulative entropy) to confirm the inference of causal direction exhibited by the direction between different nodes in the tree structures obtained earlier.

Conclusion

In conclusion, I discussed the various state of the art approaches to determine the causal direction in random variables. There are some assumptions and choices made for approximation in Vreeken[1] which were discussed critically above. I also discussed some of the further improvements in this area of research. Following to that, there was another interesting remark about the approach for dealing with datasets of large dimension and when many of the dimensions are assumed to be relatively unimportant. I am intrigued by the idea if this concept of determining causality could be implemented on fiction story analysis and character familial and sentiment relations to come up with better story graph representations from the same. I am also interested ahead to learn about how learning causality structure patterns could improve the existing state of the art. Causality has always been the subject of research and something to talk about. A journalist Sam Levenson once said, *"Insanity is hereditary. You can get it from your children."*

References

[1] Vreeken, J. *Causal Inference by Direction of Information.* In Proceedings of the SIAM International Conference on Data Mining (SDM'15), SIAM, 2015.

[2] Heikinheimo, H., Seppnen, J.K., Hinkkanen, E., Mannila, H. and Mielikinen, T. *Finding low-entropy sets and trees from binary data.* In Proceedings of the 13th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), San Jose, CA, pages 350-359, 2007.

[3] Kailash Budhathoki, Jilles Vreeken *Causal Inference by Stochastic Complexity* In: Proceedings of the SIAM Conference on Data Mining (SDM), SIAM, 2017

[4] Peters, J., Janzing, D. and Schlkopf, B. *Identifying Cause and Effect on Discrete Data using Additive Noise Models.* pages 597-604, 2010.

[5] Kailash Budhathoki and Jilles Vreeken *Causal Inference by Compression* Data Mining (ICDM), 2016 IEEE 16th International Conference

[6] Alexander Marx, Jilles Vreeken *Causal Inference on Multivariate Mixed-Type Data by Minimum Description Length*

[7] Emma Persson, Jenny Haggstr, Ingeborg Waernbaum and Xavier de Luna *Data-driven Algorithms for Dimension Reduction in Causal Inference* Computational Statistics and Data Analysis, Elsevier

[8] Roel Bertens, Jilles Vreeken, Arno Siebes *Keeping it Short and Simple: Summarising Complex Event Sequences with Multivariate Patterns*