Assignment 3: Interaction
Caveat Lector

Shrestha Ghosh (2567717, s8shghos@stud.uni-saarland.de)

July 4, 2018

1 Introduction

In this report we look into the principle of deriving the causal direction between two data objects by comparing the amount of information one object provides about another and vice versa. While Kolmogorov complexity provides a strong theory it is not practical to implement it. Practical implementations require heuristics to estimate the Kolmogorov complexity.

2 Direction of Information to infer causality

The goal of the work proposed by Vreeken [3] is to infer the direction of causality given two correlated data objects X and Y. The first assumption made by the authors is that of causal sufficiency which means that there is assumed to be no confounding factor causing both X and Y. This is a strong assumption because it is not always possible to determine whether all determining factors have been accounted for in a dataset.

While no assumption is made on the distribution and the domain of the data, *i.e.,* X and Y are univariate or multi-variate and have any arbitrary distribution, X and Y can only take real values. The model can handle both discrete and continuous real-valued data in principle and practice. In principle, (cumulative) entropy is defined for univariate data for which we require probability distribution functions (pdfs) for conditioning variables and cumulative distribution functions (cdfs). In practice, cdfs are easily available from the observed data but we need to estimate the pdfs because we do not know anything about the distribution of the data and extend the formulations to multivariate case. The extension to multivariate case is easily achieved by employing the chain rule for joint distribution of random variables using conditional probabilities. Once we apply the chain rule we isolate each dimension conditioned when required and hence can determine the conditional entropy.

In theory direction of information can be inferred from the Kolmogorov complexity of one data object conditioned on the other. Here the authors assume that observed data, both X and Y may contain noise which can be introduced with the cause or later with the effect or both. In order to make the inference robust to noise, the Kolmogorov complexity of one data object, say Y, is developed by conditioning on the model of the other data object (*i.e.,* the compressible part of X) instead of considering the entire data. This makes sense, because trying to explain noise-distorted Y by conditioning on noise-distorted X (described by its complexity $K(Y|X)$) will take more bits than $K(Y|X')$ but will not give additional information about Y because in the former case we are modeling noise. Essentially we would like to explain observable effect Y with respect to a model X' best describing the cause X. The two part Kolmogorov complexity is required to account for noise in Y which may be introduced after the cause and is hence not explained by X'. The assumption of X' causing Y' prevents overfitting the model to the observed data. The second assumption regarding normalization is justified because both X and Y may not have same dimensions and/or ranges. However the normalization approximation is not intuitive. If we analyze the relative conditional complexity

$$\Delta_{X\to Y} = \frac{K(\mathbf{Y}'|\mathbf{X}')}{K(\mathbf{Y})} + \frac{K(\mathbf{Y}|\mathbf{X}',\mathbf{Y}')}{K(\mathbf{Y})} \tag{1}$$

and the approximation,

$$
\begin{aligned}
\Delta_{X\to Y} &= \frac{1}{2}\left(\frac{H(\mathbf{X}',\mathbf{Y}')}{H^u(\mathbf{X}',\mathbf{Y}')} + \frac{h(\mathbf{Y}|\mathbf{X}',\mathbf{Y}')}{h^u(\mathbf{Y}|\mathbf{X}',\mathbf{Y}')}\right) \\
&\geq \frac{1}{2}\left(\frac{H(\mathbf{X}',\mathbf{Y}')}{H^u(\mathbf{X}')+H^u(\mathbf{Y}')} + \frac{h(\mathbf{Y}|\mathbf{X}',\mathbf{Y}')}{h^u(\mathbf{Y})}\right) \qquad (2)
\end{aligned}
$$

where Eq. (2) is obtained after applying the upper bounds and equality holds for independence. Thus independence assumption leads to lower estimation of $\Delta_{X\to Y}$. Also, in principle, unconditional models X' and Y' remain same in both directions, $\Delta_{X\to Y}$ and $\Delta_{Y\to X}$, which means that we are effectively using only the second term in the difference of relative amount of directed information for inferring direction of information. Approximation is more aligned to Eq. (1) if we consider,

$$
\Delta_{X\to Y} = \frac{1}{2}\left(\frac{H(\mathbf{X}',\mathbf{Y}')}{h^u(\mathbf{Y})} + \frac{h(\mathbf{Y}|\mathbf{X}',\mathbf{Y}')}{h^u(\mathbf{Y})}\right) \qquad (3)
$$

where normalization is w.r.t the entropy information of the observed effect data object.

So far we have discussed how the data is described and how to approximate Kolmogorov complexity using cumulative conditional entropy. Now we shall see how entropy approximations is done in practice. Unconditional entropy for a data object is calculated by greedily selecting the dimension with lowest entropy conditioned on previous dimensions and then discretizing it. This is a faster heuristic to get the factorization approximation for a multivariate data object. The authors present a dynamic programming optimization to estimate the conditional cumulative entropy. Selecting the minimum entropy of a data object $H(\mathbf{Y}|\cdot)$ over all factorizations of the dimensions of $\mathbf{Y}$ is important, because in real data attributes are not always independent of each other. In the equation (4) approximating the conditional cumulative entropy of a data object $\mathbf{Y}$ conditioned on discretized data object $\mathbf{X}'$, when we get the dimension with lowest entropy (here $\mathbf{Y}_{\sigma_Y(i)}$) at any iteration i, we are bound to catch, if present, any dimension correlated with or multi-collinear to the conditioned attributes. This helps us to estimate a tighter bound. For example if $\mathbf{Y}_{\sigma_Y(2)}$ is dependent on $\mathbf{Y}_{\sigma_Y(1)}$ the entropy of the second dimension will be much less when conditioned on the first.

$$
h(\mathbf{Y}|\mathbf{X}') = \min_{\sigma_Y} h(\mathbf{Y}_{\sigma_Y(1)}|\mathbf{X}') + h(\mathbf{Y}_{\sigma_Y(2)}|\mathbf{X}',\mathbf{Y}'_{\sigma_Y(1)}) + \ldots + h(\mathbf{Y}_{\sigma_Y(l)}|\mathbf{X}',\mathbf{Y}'_{\sigma_Y(1)},\ldots,\mathbf{Y}'_{\sigma_Y(l-1)}) \qquad (4)
$$

The experiment settings considered for causal discovery in real world data can be improved. In the work the authors first mine for correlated dimension pairs and then determine the direction of causality within each pair. Firstly in this process the causal sufficiency assumption may no longer hold. We may get (X_a, X_b) as one correlated pair and (X_a, X_c) as another. Secondly, this method limits the scope of ERGO which can deal multivariate data. The task of deriving correlated data objects from a dataset in itself is a hard task. For a dataset of n dimensions, we can select two data objects in $\binom{n}{l}\binom{n-l}{k}$ ways where l and k are the dimensions of the two data objects, $l \geq 1$, $k \geq 1$ and $l + k \leq n$. However we can create multivariate data objects using the correlated pairs. For *e.g.,* if we have correlated pairs (X_1, X_3), (X_1, X_4) and (X_4, X_2), we can consider data objects $X = \{X_1, X_2\}$ and $Y = \{X_3, X_4\}$.

3 Inferring causality from trees

We saw how Vreeken [3] propose to infer causality for real-valued data. In this section we analyze the work of Heikinheimo et al. [1] where subsets of attributes having low entropy in a binary dataset are mined. While this work concentrates on pattern discovery it can be used to understand causal relations present in the data. Frequency based pattern mining allow itemsets where items occur together, however, skewed distributions from locally interesting patterns cannot be found. The authors argue that low entropy itemsets can help us identify these local patterns. The explanation is intuitive - if an itemset has low entropy value this means that the projection of data transactions on this itemset is more likely. Entropy increases for more unlikely itemsets. Knowing the locally interesting patterns in itself is not interpretable - how do the items interact that makes them interesting? This is answered by imposing a tree structure on the itemsets. Considering tree patterns instead of itemsets can actually help infer the direction of information. Unlike in ERGO, we do not have to define two data objects as cause and affect candidates since the parent child relation in the tree is indicative of the cause and effect.

The trees considered here have two semantics because of the restriction that an attribute cannot occur more than once in a tree (no diamond structure). Consider the equation for calculating entropy of a D-tree Eq. (5) and a U-tree Eq. (6) for a tree T with root attribute A and $T_1, \ldots, T_k$ being the k child nodes of T with $A_1, \ldots, A_k$ being the attributes of the respective child nodes.

$$H_D(T) = H(A|parent(T)) + \sum_{j=1}^{k} H_D(T_j) \tag{5}$$

$$H_U(T) = H(A|A_1, \ldots, A_k) + \sum_{j=1}^{k} H_U(T_j) \tag{6}$$

If we recall the conditional entropy equation, Eq. (4), defined in [3], we can draw parallels with D-trees and U-trees. For the D-tree, the entropy is minimized for child nodes (analogous to effect Y) conditioned on the parent (analogous to effect X). The child nodes are independent of each other and each child has only one parent. At each level of the tree, a parent and its children give the direct causal relationship. However, the cause has only one dimension (each node has a single attribute). For multivariate cause we can turn to U-trees - each node is conditioned jointly on its child attributes. Conversely, in U-trees the effect is univariate. But, since the model mines all trees satisfying the frequency and entropy thresholds, we could find multivariate cause when the model returns two D-trees with different roots but same child nodes and similarly for U-trees. We should also note that the tree models the entropy of data in its entirety including any noise that may be present in the data. Moreover, the entropy is calculated using the relative frequency of the projection of data transactions on an itemset. As we know from the previous section, pdfs cannot be estimated from observed data hence the entropy calculations done by this model may not reflect the true distribution of the data.

Besides, setting a threshold for tree entropy, the model also regulates the branching factor of the tree to a very low number (two or three). Another way to adapt the model for causal inference would be to limit the height of the tree to one and not regulate the branching factor such that the model returns the low entropy trees. This would be more comparable to ERGO which looks into direction of information without confounder.

The authors also explain how to mine high entropy sets in order to find interesting patterns. While high-entropy sets make interesting patterns - they can help detect anomalies, outliers, something away from the normal distribution in the data - they are not interesting in causality point of view.

4 How can we improve?

Both ERGO and low-entropy trees depend on computing conditional and unconditional entropy estimates of attributes. Because they are estimated there is a chance of inferring wrong conclusions regarding directionality because we are not sure of the margin of error. For example, take the relative conditional complexities $\Delta_{X\to Y}$ and $\Delta_{Y\to X}$ values for real world data computed by ERGO. We know that lower the value, stronger is the causality in that direction, but, how significant is the difference in the direction relative information? And, how significantly lower than 1 are the $\Delta_{(\cdot)}$ values? Moreover, if the difference is really small we are unable to determine if the relation is actually a bijection. Marx and Vreeken [2] model causal inference using minimum description length (MDL) on any type of data (univariate, multivariate, discrete, continuous and binary). The MDL of the causal model and the data given the model is applicable to all causal models and is therefore uniformly comparable across models. Here the authors have proposed tree models which encode a probable effect by splitting on attributes forming the probable cause. The tree models defined here are unbiased between the gains obtained in two directions and also unbiased between individual attributes of the cause and effect variables.

References

[1] Hannes Heikinheimo, Eino Hinkkanen, Heikki Mannila, Taneli Mielikäinen, and Jouni K Seppänen. Finding low-entropy sets and trees from binary data. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 350–359. ACM, 2007.

[2] Alexander Marx and Jilles Vreeken. Causal inference on multivariate mixed-type data by minimum description length. *arXiv preprint arXiv:1702.06385*, 2017.

[3] Jilles Vreeken. Causal inference by direction of information. In *Proceedings of the 2015 SIAM International Conference on Data Mining*, pages 909–917. SIAM, 2015.