Markov Blanket Causal Discovery Using MML

kl

5 September 2017

Motivation

- Causal discovery refers to an algorithm-based automated process of reconstructing the generating causal model from observational data.
- Here is an example of a causal model, called the Alarm network. It is a medium-sized model with 37 variables, so learning back this generating model structure isn't too difficult with thousands of samples.
- But imagine a large structure contains hundreds/thousands of variables. In this case, neither exact nor approximate causal learners can finish the task within a reasonable time while still keep similar accuracy as for smaller models.
- The motivation of this project is to scale up causal discovery to large models, possibly contain hundreds of variables while still obtain state-of-the-art reconstruction accuracy.
- In early 2000, people started splitting the big problem into smaller ones and tackling them independently. This approach has been summarized as the local-to-global approach.
- To demonstrate how the LGL approach works, let's look at the Alarm network again.
- As the first step, a variable is picked as the target, whose surrounding variables are learned. The choice of a surrounding variable subset is not fixed, but there are two popular ones, which are the neighbours or the Markov blanket. This step can be repeated independently to all variables.
- The following step is to learn a substructure within each variable subset.
- The final step is to 'stitch' all the sub-structures together to obtain a global network. If some of the neighbouring sub-structures have conflicts on edge existences or directions, a consensus must be reached before the stitching process.

Progress

. . .

Markov blanket

- In this slide, we introduce the concept of MB. The MB of a variable X is the minimal subset of variables that satisifies the following conditional independent relation.
- What this equation saies is given the MB of X, X is independent from everything else.
- To visulize this concept, let's look at a Bayesian network. The MB of X in a BN consistes X's parents, children and children's other parents, which we normally call spouses.
- Here is an important feature of MB that I want to mention, the symmetry condition. If X is in the MB of Y, then Y is also in the MB of X. Imagine X is a parent of Y, then Y must be a child of X. And similarly, if X is a spouse of Y, then Y is also a spouse of X. The symmetry condition holds for all roles in a MB, so it also holds for the entire MB.
- This is an important feature that has to be satisfied during the local-step, because if the learned MB of X contains Y and Z but the learned MB of Y doesn't contain X, then there is a problem.

Minimum message length

- Minimum message length, MML, is a scoring metric that we used throughout this project for the purpose of MB discovery, and both local and global structure learning.
- It was developed by C. W. in the 1960s as a way to balance the complexity of a statistical model and the fit of this model to a given dataset.
- MML is a two-part message length that is commonly written down in this form, where the 1st part is the message length of a hypothesis/model, and the 2nd part is the message length of data given the hypothesis.
- Since it is infeasible to calculate the strict MML in most cases, Wallace and Freeman proposed an approximation to the strict MML, which is known as MML87.
- The first term in I(H) is the message length for using the parameters θ, and the last two terms depend upon the precision of these parameters.
- In the 2nd part, we have a negative log likelihood and the cost of using limited parameter precision.
- In general, most of these terms aren't too difficult to calculate, except for the Fisher, which measures the sensitivity of the neg log likelihood to different parameter values.
- The Fisher is the determinant of the FIM, which is the expected 2nd deriv of the neg log likelihood. It isn't a pleasant thing to directly calculate the fisher for some models, such as the logit and NB models that you will see in a minute.

Markov blanket discovery

- The problem of learning the MB candidates of a varaible is call MB discovery.
- There are two major approaches in MB disc, the first approach uses CI-test to assess whether or not a variable belongs to the MB. This approach normally starts with empty set and gradually include candidates along its way. At the end of the inclusion step, there is a process to delete any false positives.
- The second approach learns local structures around the target and reads off the MB candidates from the learned sub-structure.
- The way we tackle this problem is similar as the first approach but not from a CI point of view.
- The definition of MB makes it the smallest informative variable subset for a target. Therefore, we can use MML with an appropriate predictive model to search for such a subset.
- The models we used are . . .

MB Disc using MML FOM

- The FOM is a simple model by omitting the higher order interaction terms. In other words, it assumes the input variables' predictive power for the target are additive. It is a strong assumption that is unlikely to be real but captures a certain degree of information for the target. The benefit of having this assumption is to reduce the required sample size for predicting the target.
- Most work for FOM has been done in the 1st year.
- The message length of FOM can be calculated using this formula.
- Use MLE estimation instead.
- Because the complexity of the Fisher, it is inefficient to calculate the msg len of MML FOM. And this is only for binary target variable, if we generalize this to multinomial, the calculation will be even more difficult.

MB Disc using MML CPT

- Quite a different model we used is the conditional probability table.
- A CPT models the marginal distribution of the target w.r.t. the input variables. It can represent any distribution provided there are enough data, because a CPT pretends all inputs are parents of the target variable.
- For each parent's instantiation, the target variable is a multi-state distribution, whose message length can be calculated using the multi-state MML that was developed by wallace and boulton in 1968.
- Being a discrete model, the multi-state MML can be efficiently calculated using the adaptive code approach which doesn't directly calculate the Fisher.
- The total message length of the target given the inputs is the sum of ms-MML over all parent's instantiations.
- The last term in this formula is the constant difference b/w the adaptive code and MML87 approaches, because MML87 states the MML estimation of the parameters, while the adaptive code approach doesn't.
- Because a CPT pretentds the inputs are parents of the target, it has exponential number of params in the number of parents.
- This makes a CPT model requires more data for effective search.

MB disc using MML NB

- While a CPT express the full expressive power of a parent set, it could cause a search to fail due to excessive 1st part of MML.
- A NB model requires fewer parameters than a CPT by assuming all inputs are children of the target with the conditional independence assumption.
- Same as the FOM case, direct calculation of NB's message length using MML87 is quite inefficient and unstable due to the complex Fisher. The detail of its Fisher is written in the report.
- Hence we developed a way to use the adaptive code approach to calculate its msg len.
- We started working on this problem only recently, so I don't have a closed form formula for the adaptive code MML, but the idea is. . .
- Explain an example, predict sex of a person based on height, weight, foot size

The MBMML algorithm

- Go through this algorithm

Experiments on MBMML+CPT's accuracy

- In this slide, there are two comparison results b/w MBMML and SLL and HITON.
- SLL is a LGL causal disc algorithm that intend to find the true model structure by using dynamic programming and the BDe metric. Its returned MB candidates are read off from the learned sub-structure around the target.
- HITON is a MB disc algorithm uses CI-test for variable inclusion.
- The experiments were conducted on 11 real models with 3 different sample sizes. For each sample size, the learning was repeated on 10 different datasets in order to report an averaged performance of these methods with 95 per cent confidence intervals.
- The reported statistic is the edit dist, which is defined as the minimum number of operations required to get from the learned to the true MB. Example. . .

Bayesian model averaging

- So far, I have presented three different models for MB disc using MML. The CPT and NB treat input variables differntly from each other, the cpt pretends . . . whilst the NB pretends . . .
- But it is arbitrary to fix a model str because we don't know the true model.

Markov blanket polytree

- A polytree is a . . .
- We define a MBPT as . . . , in other words, except the node Y, every other nodes in P can only be Y's parents, children and/or spouses.
- The motivation of learning polytrees within feature subsets is to reduce the complexity of local learning.
- As subgraphs of DAGs, the space of polytrees is much smaller than DAGs as we can see from this table. For a modern computer, it's not too bad to search through the entire space of polytrees with 7 nodes in the MB.
- Although we limited the local structures to be singly connected, we conjecture polytrees provide sufficient connections for learning a global causal model.

Enumerating all MBPTs

- In this slide, we introduce a way of calculating the number of MBPTs for a given MB size.
- The proof of this proporsition also sketched a systematic way of enumerating all MBPTs.
- Once the space is full enumerated, we can search through the entire space to find the optimal local structure w.r.t. a scoring function.
- The MML CPT metric is a decomposable score that can be used as an objective function for search for the optimal MBPT.
- The decomposibility of this metric makes it efficient to calculate the message length of a structure by summing over the msg len of each node given its parents.

Deterministic global step

- Before ‘stitching’ the sub-structures into a global structure, there needs to be a process to check for any feature conflicts.
- The feature here refers to either edge or arc.
- If there is an edge conflict b/w two sub-strs, a deterministic global step forces the one contains the edge to drop it, or the one doesn’t contain the edge to add it.
- If there is a direction conflict b/w two sub-strs, a deterministic global step removes the direction.

A local-to-global causal disc algorithm

Global str learning results

Probabilistic global step

- Unlike the deterministic step, the prob step incorporates feature confidence into a/c.
- These confidences can be estimated using bootstrapping and repeat the same learning process on each bootstrapped sample. The confidence of a feature is measured by the number of occurance of this feature out of the total number of bootstrapped samples.
- CaMML is a causal learner based on mml and mcmc. It takes various str priors to help accurately sampling its search space.
- By giving these feature confidences to camml as arc prior, it helps reducing camml's sampling space and hence increase its scalabilty to deal with larger models.
- Currently, it stops working at about 50 variables.
- We've done some initial experiments on helping camml's scalability by giving arc priors, but the results are not as we expected. Some results with priors are even worse than with no priors at all.
- We suspect this is due to several reasons . . .

Future work

- Bayesian model averaging . . .
- Scale up CaMML . . .
- An intelligent system under which each local structure learning is able to communicate with its neighbours on its findings.
- By starting with the subsets of variables with high confidence, their findings can be passed to its neighbours as structure priors.
- At the end of the local learning process there should be fewer or even no disagreements on adjacencies.