Distinguishing Cause and Effect: Inferring causal direction in two variables

Lohit Jain[1]
Balram Meena[1]
Advisor: Dr. Amitabha Mukerjee[1]
{lohit,balram,amit}@iitk.ac.in

[1] *Dept. Of Computer Science and Engineering, IIT Kanpur*

April 25, 2014

Abstract

The task of attributing cause and effect is prevalent and pervasive in history as well as present everyday lives. The factors affecting economy, human health, global warming can help us in solving many problems. The standard way to detect causal direction is to perform random experiments, which can be expensive, unethical or even impossible to perform. Inferring the same from already collected data can solve many problems. In this project we dealt with finding causal direction among two variables, given their distribution. This problem was part of the Cause Effect pair challenge. We used functional noise model and deterministic relational models to find the cost involved in causal directions for the two variables . We determined the existence of a relation using a SVM classifier on features extracted from the distribution of data.

1 Introduction

Causality inference is an important task in many fields of life. Causal inference in medicine, physics, economics, geophysics etc have important applications. Determination of factors affecting our health, climate etc can enable us to improve them in a meaningful way. But the gold standard for inferring causality is to perform random experiments. These experiments can be costly (example in the field of economy), even unethical (experiments of human health) or unfeasible (climate change). A possible solution is to use data from other sources. Such data is easily available from routine data collection worldwide. Say, for example we are tackling

the problem of fighting lung cancer. Performing random experiments for factors of cancer is unethical and banned. But we can use the existing medical data of people smoking with lung diseases and/or cancer. This data can be used to infer causal dependence of cancer on smoking. Such observations can help us prevent lung cancer by discouraging its causal factors like smoking.

The Cause Effect pairs challenge aimed at unraveling such potential cause effect relationships from observational data. Consider for instance, variable 'Crime reports in locality'. We need to find the factors affecting this variable, like number of Police stations. *"The objective of the challenge", as stated," was to rank pairs of variables A, B to prioritize experimental verifications of the conjecture that A causes B"* [9]. But observing correlation (different from causation) or some data dependency of A on B does not necessarily mean causality. They can be consequences of a common cause.

The problem is to find confidence of causal directions among pairs of variables given the distribution. We assume the absence on any feedback loops in the data. The data is time independent, with variables being aggregate statics. We also assume independence between two pairs of data.

We deal with this problem, inferring causal direction for pair of variable X,Y, in three steps. The first step is to determine the model costs involved in fitting a causal direction ($X \Rightarrow Y or Y \Rightarrow X$) to the data. The second step is to classify pairs as causally dependent ($X \Rightarrow Y or Y \Rightarrow X$) or independent (Independent or effects of common cause).

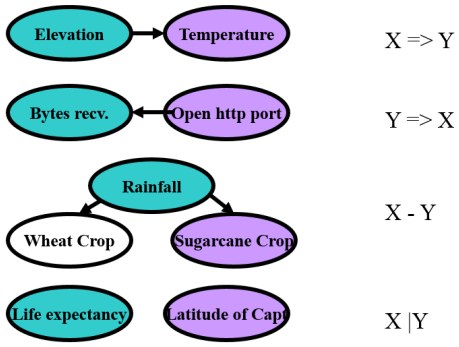

Figure 1: Possible Causal Relations between X and Y

2 Related work

Inferring causal relations for DAG's of greater than 2 variables was done by analyzing conditional in-dependencies [6] and constraint based approaches (Pearl, 2000; Spirtes et al., 1993). However such models are unable to differentiate between Markov DAG's with similar set of dependencies (and in-dependencies).

Many models infer causal directions, $X \Rightarrow YandY \Rightarrow X$, by using an independent noise term E to find the direction of lower total complexity.
Latent variable models (LiNGAM) [7] model effect as linear combination of cause and noise,$Y = pX + qE$. The model assumes that Y is linear combination of cause X and independent noise term E.
Hoyer et al. [1] introduced additive noise models, modeling the causal direction $X \Rightarrow Y$ as $Y = f(X) + E$ for the special case of two variables. The assumption is that Y is function of cause X and independent noise term E. The absence of such additive noise model in the opposite causal direction is used to infer causality.
Zhang et al. [8] proposed a post-nonlinear model, allowing a further non linear transformation on X and E are allowed, modeling effect as $Y = h(f(X) + E)$ with E statistically independent from X. A hetero-schodastic noise model was discussed which assumed $Y = f(X) + E.g(X)$.
Mooij et al. [3] considered hypothetical effect variable to be a function of the hypothetical cause variable and the independent noise term, $Y = f(X, E)$. The priors on f were assumed to be general non parametric priors. The causal direction was inferred using standard Bayesian model selection.

Although in general observational data has reasonable errors associated with them, but for high accuracy data the added assumption of noise will not hold. Janzing et al. [4] assumed a deterministic relation between X and Y, $Y = f(X)$. They considered the non trivial case of f being a invertible function. This model is shown to be linked to information geometry.

All of the mentioned models assume existence of causal directions (either $X \Rightarrow Y$ or $Y \Rightarrow X$). This is not true for real observational data. When investigating causal directions in observational data, we can not be sure of existence of direct causality. Example, average humidity and pollution levels are not directly related causally.

3 Dataset and Evaluation Scheme

We used pair distribution data provided in the official competition. The Data is taken from casualty workbench. Data contains hundreds of pairs of real data

variables pairs with known causal relationships from varied domains (chemistry, climatology, medicine, physics, sociology, ecology, economy, engineering, epidemiology etc).
The pairs consists of independent variables, dependent variables (but not causally related) and causally related variables intermixed. Some semi-artificial cause-effect pairs (Given real variables mixed in several different ways to produce a given result/outcome) are also added.
We have used 1200 pairs of variables (mix of real and simulated) as training set. We use 1000 pair of variables (real observational data) as test set.

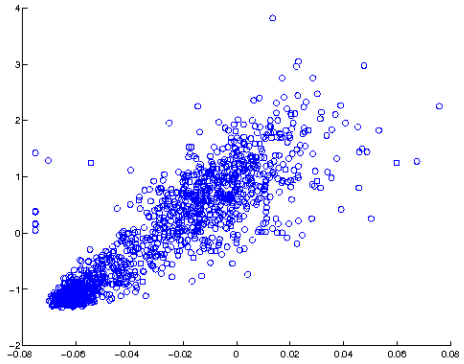

Figure 2: Distribution between Bytes sent and Open http connections at a moment

For each pair of variables we produce a score between +Inf and -Inf. Positive values indicate higher certainity of the causal direction $X \Rightarrow Y$, whereas large negative values $Y \Rightarrow X$. Values close to zero indicate absence of definite causal direction, with score 0 implying no causal relation $(X - Y)$ at all.
We evalueate two Area Under the ROC Curve scores, one for success in detecting direction $X \Rightarrow Y$ and other $Y \Rightarrow X$. Due to the symmetry of the problem, we take the average as our final score.

4 Methodology

The methodology is divided into three steps. In the first step we compute the cost and minimum description complexity involved with the causal direction $X \Rightarrow Y$ and $Y \Rightarrow X$, in functional noise and deterministic models. Functional noise model assumes hyper priors on evidence of $Y \Rightarrow X, Y = f(X, E)$ and uses standard Bayesian model to find the correct causal direction. Deterministic model assumes $Y = f(X)$, with f an invertible function. The model uses the fact that

independent selection of f and probability will make the effect Y dependent on f in some sense.

In the second step we extract features from the distribution data. These features are used to train a classifier to predict the existence of causal dependency among X and Y.

In the last step, we combine the above two steps. If the classifier determines existence of a causal direction, we score the pair according to cost generated in the first step. If the pair are classified as independent, score is set to 0 to reflect absence of causal direction.

4.1 Cost of Models

We compute the cost of fitting the distribution over following two models:

4.1.1 Gaussian Process Inference (GPI)

Gaussian processes (GPs) are used for Bayesian non-parametric estimation of latent functions. It approximates Gaussian processes for regression and binary classification. The Gaussian prior is defined by providing its mean and covariance. Since in our model we used unsupervised learning, the negative log marginal likelihood (with partial derivatives relative to hyper-parameters) is computed. This is used to fit the hyper-parameters to the given data (utilizing BFGS minimization algorithm for parameter estimation). The normalized residuals are considered to be initial values of E. These values favor functions and distributions of lower complexity.

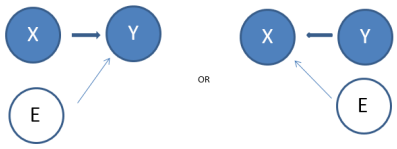

Figure 3: GPI model: $Y = f(X, E)$ or $X = f(Y, E)$

This model makes the following additional assumptions [3]:

1. *Y is an effect of only X and E. $Y = f(X, E)$, where f is the causal mechanism.*

2. *X and E are independent and have no common factors.*

3. *Distribution of* $\{X, Y\}$ *is not dependent on f.*

4. *Priors to noise have form :* $E \sim N(0, 1)$.

For cost involved in direction $X \Rightarrow Y$, this model involves solving the following marginal likelihood equation [3]. Here x_i, e_i and y_i are values of X, E and Y, f is causal mechanism, N number of distribution points δ is Dirac-delta function and θ_P is the prior for corresponding P.

$$p(x,y) = [\int (\prod_{i=1}^{N} p(x_i|\theta_x))p(\theta_x)d\theta_x][\int (\prod_{i=1}^{N} \delta(y_i - f(x_i, e_i))p_E(e_i))de * p(f|\theta_f)df * p(\theta_f)d\theta_f]$$

4.1.2 Information Geometric Causality Inference (IGCI)

This is a standard information geometric model to find causality in low noise variables. It involves logarithms of step distributions of the values among the two variables. We use the score from this model directly as a metric for causal direction.

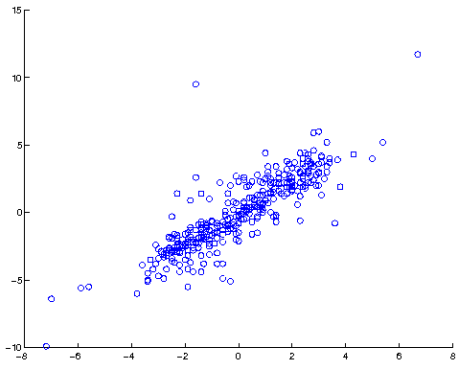

Figure 4: Distribution between average annual rate of change of population and total dietary consumption for (kcal/day). The results in first step: 0.9271. X⇒Y

4.2 Dependency Classification

Experiments with gpi and igci generated poor results in the cases where no causal relation existed between the two variables. This is because both models assume the existence of a causal direction and then try to find the correct direction. Example in figure 5, there is no causal direction between female expectancy and

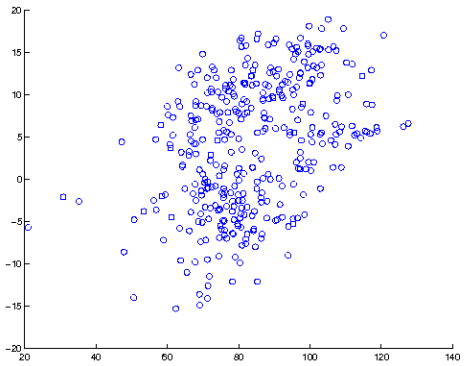

Figure 5: Distribution between life expectancy at birth for different countries, female, 2000-2005 and latitude of the country's capital. First step score: -0.241. Y⇒X. Wrong!

latitude of a country's capital, but our first step returns a causal direction. To counter the fitting nature of gpi and igci models, we chose features that could be useful in classification of pairs into independent or dependent (causally).
For this purpose we considered statistical features and selected the following features for our purpose:

- **Gini-coefficient**: This is a measure of *statistical dispersion*, measuring the inequality among values of pair value distributions. Zero implies perfect equality whereas 1 implies maximum inequality. But practically, for larger distributions, extremeties are less likely [11].

- **Normalized HSIC**: Hilbert-Schmidt Independence Criteria is an statistical independence criterion based on the eigen-spectrum of covariance operators in reproducing kernel Hilbert spaces.

- **Normalized entropy**: Measures the uncertainty of the conditional distribution for both causal directions.

- **Kurtosis**: Measure of the peakedness nature of the distribution of X and Y.

- **Skewness**:Measure of the asymmetry of the distribution of a wrt. mean value. This can be non zero or undefined.

- **Divergence**: Divergence expresses the distance of conditional distributions for both causal directions.

- **Count**: Count of X and Y, equal in our case.
- **Unique count**: Unique occurrences in the distribution of X and Y.
- **Fit and Fit error**: The error and cost relating to fitting polynomial distributions over the distribution of data in X and Y.
- **Joint entropy**: Measure of uncertainty attached to X,Y.
- **Discrete mutual information**: Similar to transformation between two variables. This is measure of mutual dependence of two variables.
- **Adjusted mutual information**: Measure of comparing clustering of variables (relates to variation of information among the pair).

We trained a SVM classifier using linear kernel on training set to classify pairs into causally related and unrelated, using the above features.

4.3 Combination of results

To combine the results we give priority to classifier scores. If the classifier classifies pair as causally independent, we return the score for this pair as 0. For the other case, we combine the results of GPI and IGCI, maintaining the sign of the values.
If both have same signs, $FinalScore = gpi * igci$, otherwise $FinalScore = (gpi - igci)^2$.

5 Results

The final Area under the ROC score was found to be **0.658**. This easily surpassed the python benchmark of **0.570** by the challenge authorities. The results solely based on IGCI model coupled with classifier got the Area Under ROC score **0.597**.

The highest score in the cause effect challenge was **0.819** followed closely by **0.810** and **0.807** [12]. The huge difference between the scores can be attributed to higher training data used and large number of extracted features (around 1000) from the distribution. The cost estimation in submission of highest scorer, was also done through supervised learning.

Even though our method is outperformed, our method is very efficient in terms of computation cost and resources. Due to the unsupervised nature of cost estimation, the code can be easily distributed over multiple machines.

The individual results with GPI and IGCI models were less than that of the combination. This is due to the limitations involved in the two models. GPI limits effect to function of cause and noise, where as IGCI assumes absence of noise. In normal observational data, the existence independent data (noise) is restricted in some cases, but is prominent in other.
The combination of GPI and IGCI produced accurate predictions in the cases where X and Y were causally related ($X \Rightarrow Y$ or $Y \Rightarrow X$), but inaccurate results for cases of independence (X and Y both independent or result of common cause).

The accuracy of the classifier was about 73%. The features selected are efficient for the task of determining existence of direct causal relation.

6 Conclusions

GPI and IGCI models assume existence of direct causal relation and apply additional assumptions of functional noise or determinism on the cause-effect pair to find lower direction of complexity. The combination of GPI and IGCI performs better for real life data. For detecting cases where no direct causal relation exists other features can be utilized. We proposed a set of features, on which an SVM classifier was able to classify pairs as either causally dependent or independent. The proposed method was able to outperform benchmark but was outperformed by a more feature orientated approach.

References

[1] Hoyer, Patrik O., et al. "Nonlinear causal discovery with additive noise models." NIPS. Vol. 21. 2008.

[2] Peters, Jonas, Dominik Janzing, and Bernhard Scholkopf. "Causal inference on discrete data using additive noise models." Pattern Analysis and Machine Intelligence, IEEE Transactions on 33.12 (2011): 2436-2450.

[3] Mooij, Joris M., et al. "Probabilistic latent variable models for distinguishing between cause and effect." NIPS. 2010.

[4] Daniusis, Povilas, et al. "Inferring deterministic causal relations." arXiv preprint arXiv:1203.3475 (2012).

[5] Janzing, Dominik, et al. "Information-geometric approach to inferring causal directions." Artificial Intelligence 182 (2012): 1-31.

[6] Tian, Jin, and Judea Pearl. "Causal discovery from changes." Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence. Morgan Kaufmann Publishers Inc., 2001.

[7] Shimizu, Shohei, et al. "A linear non-Gaussian acyclic model for causal discovery." The Journal of Machine Learning Research 7 (2006): 2003-2030.

[8] Zhang, Kun, and Aapo Hyvrinen. "On the identifiability of the post-nonlinear causal model." Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence. AUAI Press, 2009.

[9] CauseEffect: Tasks - Causality Workbench. 2014. CauseEffect: Tasks - Causality Workbench. [ONLINE] Available at: http://www.causality.inf.ethz.ch/cause-effect.php?page=tasks. [Accessed 24 April 2014].

[10] Gretton, Arthur, et al. "Measuring statistical dependence with Hilbert-Schmidt norms." Algorithmic learning theory. Springer Berlin Heidelberg, 2005.

[11] Gini coefficient - Wikipedia, the free encyclopedia. 2014. Gini coefficient - Wikipedia, the free encyclopedia. [ONLINE] Available at: https://en.wikipedia.org/wiki/Gini_coefficient. [Accessed 24 April 2014].

[12] NIPS 2013 Workshop on Causality. 2014. NIPS 2013 Workshop on Causality. [ONLINE] Available at: http://clopinet.com/isabelle/Projects/NIPS2013/. [Accessed 24 April 2014].