Training Gradient Boosting Machines using Curve-fitting and Information-theoretic Features for Causal Direction Detection

Spyridon Samothrakis SSAMOT@ESSEX.AC.UK

Diego Perez DPEREZ@ESSEX.AC.UK

Simon Lucas SML@ESSEX.AC.UK

Wivenhoe Park
Colchester
Essex CO4 3SQ, United Kingdom.

Editor: Isabelle Guyon, Dominik Janzing and Bernhard Schölkopf

Abstract

Detecting causal relationships between random variables using only matched pairs of noisy observations is a crucial problem in many scientific fields. In this paper the problem is addressed by extracting a number of features for each matched pair using a selection of curve-fitting and information theoretic features. Using these features, we train a pair of Gradient Boosting Machines whose hyperparameters we optimise using stochastic simultaneous optimistic optimisation. The results show that our method is relatively successful, gaining a 3rd place in the 2013 ChaLearn Causality Challenge, hosted by kaggle. Our method is robust enough to be used in causality detection (or as part of a more comprehensive toolkit), although we believe it might be possible to considerably improve the quality of results by adding more features in the same vein.

Keywords: Causality Detection, Gradient Boosting Machine, StoSOO.

1. Introduction

Humanity's effort to understand causality and its relationship to knowledge can be observed in almost every academic field, including philosophy (e.g. Falcon (2012):"we think we have knowledge of a thing only when we have grasped its cause" quoting Aristotle) or Anthropology (e.g. see the ability for associative thinking in Frazer (1936)). Intuitively, one can think of two random variables A, B having a causal relationship in case somehow one can force an event to take place from random variable A, certain events take place from set B (Pearl, 2000). An example here would be measurements of the angle of gas pedal depression and the spinning of a wheel on a car. If one starts spinning the wheel, no movement should be detected in the gas pedal, but not vice-versa. In other words if an agent is able to effectively control a process given the possibility of doing so, we can claim that the agent's actions are causal to the states of the process.

It is not always easy to perform the controlled process mentioned above, but might happen that we have a number of observations of actions that do not involve agent actions. This in effect turns our problem into one of prediction. Does the knowledge of variable A help me predict variable B and/or the reverse? Obviously, assuming the mapping between A and B is stochastic, there can be no proof of causal direction in the sense given in the previous paragraph. We could however try to infer the causal direction if we assume some sensible set of priors over the mapping function, taking a view reminiscent of the work of Kolmogorov, i.e. assuming nature prefers simple mechanisms. In this paper we try to infer causal structure using a machine learning approach on A and B, aided by heaving feature extraction.

The rest of the paper is organised as follows: In Section 2 we present the method we used for inferring causal direction. In Section 3 we present some experimental results and analyse the resulting classifiers. We conclude with a short discussion in Section 4.

2. Methodology

There are some core concepts behind the methodology followed. Firstly, we are trying to find whether the mapping $F_1 : A \rightarrow B$ is more probable than $F_2 : B \rightarrow A$. This can be captured by trying to fit different classifiers at each direction of the data. This implicitly assumes that machine learning classifiers tend to prefer simpler models. The second concept is that information theoretic features about the data should be able to capture some of the characteristics of the underlying distributions, thus helping our overall classification task.

2.1 Data and Data Pre-processing

Our data source was the union of all samples provided by the "ChaLearn Causality Challenge"[1]. The amount of data provided is doubled by reversing all the examples given. The total number of labelled data is 32399 samples. Each labelled sample belongs to either class 1 (A causes B), class -1 (B causes A) or class 0 (where respectively the events are independent, influenced by a third cause or we cannot tell). The type of variable in each data sample is also known (i.e. categorical, binary or continuous).

2.2 Feature Extraction

What follows is a brief exposition of the features used.

1. *Spearman* ρ: The correlation coefficient ρ.
2. *Number of Unique Samples* A: Number of unique samples of variable A.
3. *Number of Unique Samples* B: Number of unique samples of variable B.
4. *Noise Independence* $A \rightarrow B$ *(trees)*: The mutual information of an additive noise model (Hoyer et al., 2008). Uses k-means++ (Arthur and Vassilvitskii (2007)) to discretise noise. Modelling is performed using Regression or Decision Trees.
5. *Noise Independence* $B \rightarrow A$ *(trees)*: As in feature 4, but trying to predict A using B.

1. The data can be found here: `http://www.kaggle.com/c/cause-effect-pairs/data`

6. *Noise Independence $A \to B$ (SVM)*: As in feature 4, with a support vector classifier or regressor as the modelling function.

7. *Noise Independence $B \to A$ (SVM)*: As in feature 5, but trying to predict A using B.

8. *Noise Independence $A \to B$ (trees) - Spearman*: As in feature 4 but, instead of mutual information, using Spearman ρ as an independence test.

9. *Noise Independence $B \to A$ (trees) - Spearman*: As in feature 5, but trying to predict A using B.

10. *Entropy A*: Entropy of Variable A. If the variable is continuous, k-means is performed and distance is measured from closest centre as a method for discretisation.

11. *Entropy B*: Entropy of Variable B. Same discretisation method as with feature 10.

12. *Uncertainty Coefficient $A \to B$* One-directional Uncertainty Coefficient. In case of continuous variables, the k-means trick from feature 10 is used.

13. *Uncertainty Coefficient $B \to A$* One-directional Uncertainty Coefficient. In case of continuous variables, the k-means trick from feature 10 is used.

14. *Predicts $A \to B$ (trees)*: Fraction of correctly classified examples or R^2, depending on whether B is categorical or continuous. A decision tree regressor or a decision tree classifier is used in all tree examples.

15. *Predicts $B \to A$ (trees)*: As in feature 14, but trying to predict A using B.

16. *Predicts $U \to B$ (trees)*: Predict B using just random variables that come from a distribution as close to A as possible.

17. *Predicts $U \to A$ (trees)*: As above but with reversed direction.

18. *Predicts $A \to B$ (SVM)*: Exactly as in the case of 14, but this time with support vector machines.

19. *Predicts $B \to A$ (SVM)*: See above.

20. *Predicts $U \to B$ (SVM)*: See above.

21. *Predicts $U \to A$ (SVM)*: See above.

22. *Uniform Symmetrised Divergence A*: Symmetrised KL Divergence between A and the Uniform distribution. As usual, discretisation is performed using k-means.

23. *Uniform Symmetrised Divergence B*: Symmetrised KL Divergence between B and the Uniform distribution.

24. *KL Divergence from Normal A*: KL Divergence of A from the normal distribution.

25. *KL Divergence from Normal B*: KL Divergence of B from the normal distribution.

26. *KL Divergence from Uniform A*: KL Divergence of A from the uniform distribution.

27. *KL Divergence from Uniform B*: KL Divergence of B from the uniform distribution.

28. *LiNGAM*: The LiNGAM causality coefficient (Shimizu et al., 2006), implemented by the original author of this method.

29. *ICGI - Normal Integration*: IGCI Gaussian-Integration coefficient (Daniusis et al., 2010), implemented by the original authors of this method.

30. *ICGI - Uniform Integration*: IGCI Uniform-Integration coefficient, as above.

All these features have been made publicly available[2].

2.3 Classifier

Two Gradient Boosting Machines (GBM see Friedman (2001)) have been used, with 3000 trees at each one. Each GBM has a learning rate of approximately 0.0063640 and a subsample value of 0.6. The minimum samples required for each tree split is 5. The first GBM_1 is trained using only samples from the class 1 versus everything else, where everything else forms class 0. The other GBM_{-1} is trained using samples of class -1 versus everything else. P_{GBM} is used to denote the probability of a sample belonging to a specific class. The score of each sample is set to $S = P_{GBM_1}(1) - P_{GBM_{-1}}(-1)$. In other words the score attributed to each sample is the probability of having causal direction from A to B minus the probability of having causal direction from B to A.

2.4 Hyperparameter Optimisation

A modified version of Stochastic Simultaneous Optimistic Optimisation (Valko et al., 2013) (StoSOO) was used to optimise the learning rate and the subsample percentage (i.e. the samples to be used in a bagging-like procedure) for the two GBMs. Intuitively, learning rate controls how fast each GBM should learn at each iteration, while subsample represents the proportion of samples (drawn without replacement) that are to be used in each iteration. Both hyperparameters affect regularisation and have tremendous impact on the ability of the GBMs to correctly classify new instances. StoSOO is a tree-like algorithm that samples the hyperparameter space by iteratively splitting it into smaller segments, which it then samples, until some cut-off point.

3. Experiments & Analysis

The resulting classifier and meta-optimisation technique is analysed in this section. Note that the Area Under the Curve (AUC) score of our classifier in the Kaggle's causality challenge test set is 0.79957. This gave us the third place in the competition out of 69 participants.

2. `https://github.com/ssamot/causality/blob/master/features/feature_functions_spearman.py`

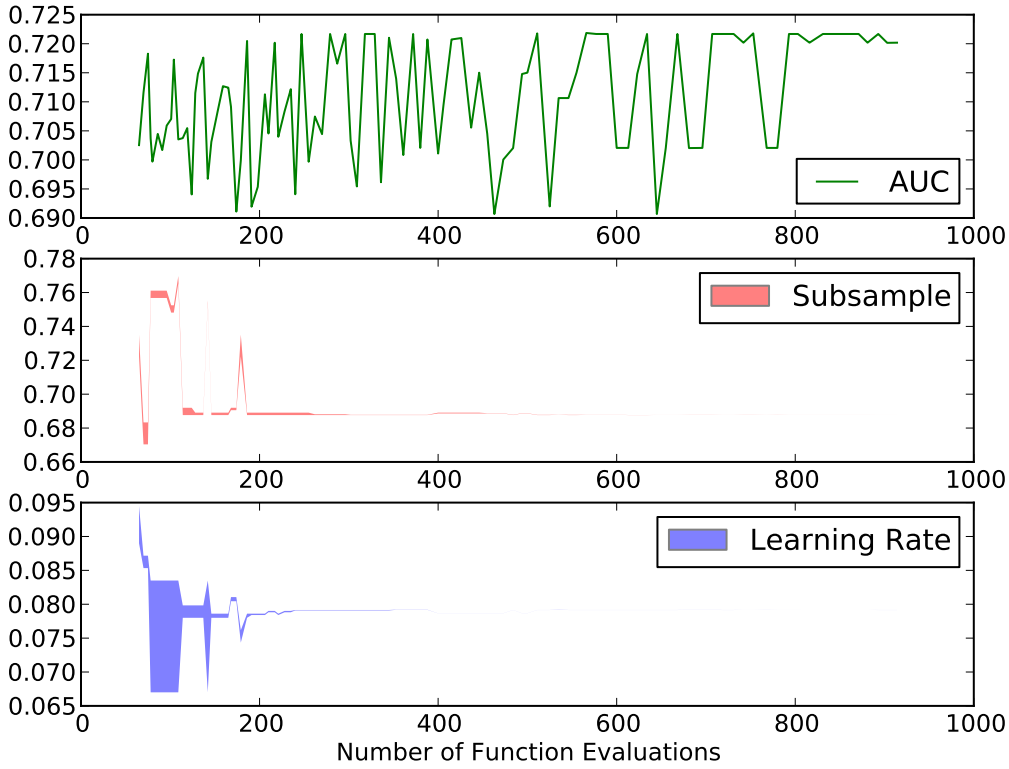


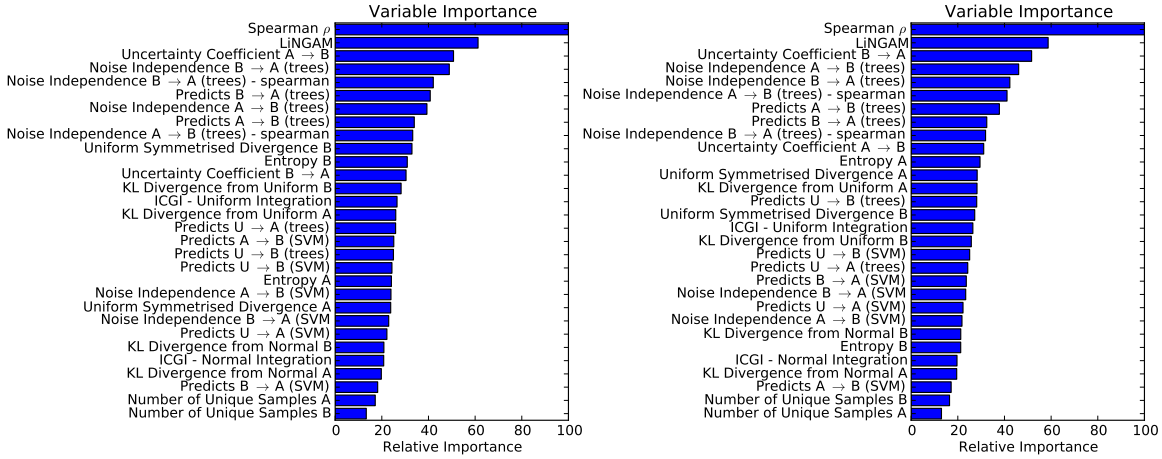
Figure 1: Hyperparameters Optimisation Progress. Notice that both hyperparameters affect regularisation. Learning rate affects speed of convergence and subsample affects the portion of samples used at each iteration of each GBP learning cycle

3.1 Hyperparameter Optimisation

A number of hyperparameters were optimized by a combination of hand-tuning and small runs of StoSOO. A sample run can be seen in Figure 1, on a subset (10%) of the experimental data, 100 trees in our GBM and a maximum tree depth of 10. Notice StoSOO improving AUC using just a subset of the data. The AUC score is obtained by doing 3-fold cross validation over a random selected subset of that data (i.e. dataset splits are NOT fixed in every iteration). Notice the randomness of AUC score (within certain bounds), but the convergence of subsample and learning rate GBM attributes. Also notice the uncertainty concerning hyperparameter values early in the run.

3.2 Training and Classifier Analysis

In Figure 2 one can see the relevant score of each variable plotted, with 100 being the score of the most important variable. Feature importance signifies the average importance of each variable, as measured by how high in the tree the variable is (being higher in the tree means affecting more samples). In an ensemble of features produced by the GBMs the normalised



(a) Feature Importances of GBM_1 (b) Feature Importances of GBM_{-1}

Figure 2: Relative feature importance for each classifier

average of these variables is what is plotted. From Figure 2 one can see that the most important feature is Spearman's correlation, presumably GBMs are first throwing away cases that are uncorrelated. The least important features involved (predictably) are the number of unique variables. LiNGAM comes second in both classifiers, which is somewhat surprising judging from the fact that the feature is a causality measurement in itself. High in importance as well is the uncertainty coefficient, which can be seen as a measure of (non-symmetric) independence. Notice that in the majority of cases the importance of asymmetric features changes depending on which GBM uses them. Finally we would like to state that we are aware that some of the features might not make much sense. The ensemble of all features however seems to provide a robust solution.

4. Conclusion

A method for detecting causality has been presented. Obvious improvements to the method include creating more curve fitting features and introducing more information theoretic features. One could, for example, add trees of different sizes, plus a number of SVMs with different kernels/kernel parameters. Fitting linear classifiers/regressors or higher level polynomials would be another option. Finally, at the beginning of this paper we emphasised the decision theoretic aspects of causality detection. It might be possible to directly tackle the problem using decision theoretic methods (e.g. standalone StoSOO or Monte Carlo Tree Search).

Acknowledgments

This work was supported by EPSRC grant EP/H048588/1 entitled:"UCT for Games and Beyond".

References

David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035. Society for Industrial and Applied Mathematics, 2007.

Povilas Daniusis, Dominik Janzing, Joris Mooij, Jakob Zscheischler, Bastian Steudel, Kun Zhang, and Bernhard Schoelkopf. Inferring deterministic causal relations. In *Proceedings of the Twenty-Sixth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-10)*, pages 143–150, Corvallis, Oregon, 2010. AUAI Press.

Andrea Falcon. Aristotle on causality. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Winter 2012 edition, 2012.

James George Frazer. *The Golden Bough: A Study in Magic and Religion. Vol. 13, Aftermath: a Supplement to the Golden Bough*. Macmillan, 1936.

Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, pages 1189–1232, 2001.

Patrik O. Hoyer, Dominik Janzing, Joris M. Mooij, Jan R. Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 689–696. 2008.

Judea Pearl. *Causality: models, reasoning and inference*, volume 29. Cambridge Univ Press, 2000.

Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, and Antti Kerminen. A linear non-gaussian acyclic model for causal discovery. *The Journal of Machine Learning Research*, 7:2003–2030, 2006.

Michal Valko, Alexandra Carpentier, and Rémi Munos. Stochastic Simultaneous Optimistic Optimization. In *30th International Conference on Machine Learning*, Atlanta, États-Unis, February 2013. URL `http://hal.inria.fr/hal-00789606`.

Appendix A. Causality challenge

Title: Training Gradient Boosting Machines using Curve-fitting and Information theoretic features for Causal Direction Detection.
Participant name, address, email and website: Spyridon Samothrakis, Diego Perez, `https://github.com/ssamot/causality`.
Task(s) solved: Kaggle Competition.
Reference: This paper.
Method: A combination of feature extraction from the sample data, Gradient boosting machines and StoSOO meta-optimisation.

- Preprocessing: Exploit Symmetries.
- Causal discovery: Gradient Boosting Machine, Curve fitting/Information theoretic features.
- Feature selection: Feature Ranking.
- Classification: Gradient Boosting Machine
- Model selection/hyperparameter selection: Cross-validation, Stochastic Simultaneous Optimistic Optimisation.

Results:

Dataset/Task	Score
Test Set	0.79957

Table 1: Result table.

- quantitative advantages: The method and ideas behind our method are relatively simple. We advocate a feature extraction strategy based on curve fitting + information theoretic features.
- qualitative advantages: There are some elements of novelty, mostly in the ideas behind extracting features and doing hyper-parameter optimisation.

Code and installation instructions can be found here: `https://github.com/ssamot/causality`