

Discovery of Causation Direction by Machine Learning Techniques



Mario de-Prado-Cumplido and Antonio Artés-Rodríguez

Universidad Carlos III de Madrid
Dpt. of Signal Processing and Communications
Av. de la Universidad 30, 28911 Leganés, Spain
{MPRADO, ANTONIO}@TSC.UC3M.ES

Abstract

In this paper, we introduce two novel algorithms that are able to direct the edges between variables, starting from a known skeleton of the causal graph. The first algorithm is a search-and-score procedure that maximizes the conditional likelihood of the graph using the BDeu score. The directions of the edges are encoded in a binary string. Then, a genetic algorithm is used to search for the optimum over the space of all possible arrangements of the directed edges.

The other machine learning technique uses classifiers to identify the “V” structures present in the graph. This constraint-based procedure takes into account chains of three adjacent nodes, $X - Y - T$. Two kNN classifiers are trained, one with Y , the other with both $\{X, Y\}$, in order to classify the node in the edge T . If the Markov Condition is satisfied, and both X and Y are relevant in the classifier, then the correct direction of the edges must be $X \rightarrow Y \leftarrow T$.

1. Motivation of the Problem

- To discover the causal relations between variables or signals is preferable or even necessary in some cases ([3], [5]).
- This paper tackles the problem of finding causal relations based uniquely in sample sets of discrete data: $\mathbf{X} = (X_1, \dots, X_n)$.
- We propose two novel methodologies to direct the edges, once the skeleton of the causal net is known:
 - A Genetic Algorithm to maximize the BDeu score.
 - A set of simple nearest neighbors classifiers to look for the collisions or “V” structures present in the net.
- The skeleton can be obtained by several methods, for example with HITON ([1]).

2. kNN Classifiers for Causal Direction

The set of relevant variables that has information about a given variable or node of the network T are within the Markov Blanket ([4]):

- direct causes,
- direct effects and
- the direct causes of the direct effects

Our algorithm searches all the chains of three nodes, $X - Y - T$, and constructs two k nearest neighbor classifiers: $\hat{T} = f_1(Y)$ and $\hat{T} = f_2(Y, X)$.

$X Y T$	$\min P_e$	conditional independence
$\circ \rightarrow \circ \rightarrow \bullet$	only Y is relevant	$X \perp T Y$
$\circ \rightarrow \circ \leftarrow \bullet$	both X and Y are relevant	$X \text{ dep } T Y$
$\circ \leftarrow \circ \rightarrow \bullet$	only Y is relevant	$X \perp T Y$
$\circ \leftarrow \circ \leftarrow \bullet$	only Y is relevant	$X \perp T Y$

Table 1: Relevant features to classify T .

There are four possibilities to orient the edges of this chain, which are listed in Table 1. If the Markov Condition holds, there is only one setting that makes X relevant, and consequently useful for the classification task: the “V” structure: $X \rightarrow Y \leftarrow T$. The algorithm is summarized in these steps:

1. Search for a chain of three variables $X - Y - T$.
2. Calculate both $P_{e,T}(Y, X)$ and $P_{e,T}(Y)$, and the associated distributions of each error probability, by means of bootstrap resampling.
3. Calculate the hypothesis test of the distribution of $P_{e,T}(Y)$ being different of $P_{e,T}(Y, X)$. If the null hypothesis (they are equal) is rejected, and $P_{e,T}(Y) < P_{e,T}(Y, X)$, then orient the edges as $X \rightarrow Y \leftarrow T$.
4. Repeat all the procedure starting from 1 until all three variable chains are processed.

The second step of the algorithm, the estimation of the distribution of $P_{e,T}(\cdot)$, is performed by resampling the training data with a standard bootstrap methodology. The error probabilities $P_{e,T}(\cdot)$ are the test error given by a 5-fold cross validation. Finally, the Cramer Von Mises two sample test ([2]) is used to prove

if the classifiers are identical. This hypothesis test computes this quantity: $T_2 = \frac{1}{4} \sum_{x_i=X_i, x_j=X_j} (S_1(x_i) - S_2(x_j))^2$, where $S_i(x)$ are the Empirical Distribution Functions of the error probabilities.

3. Genetic Algorithms for Causal Direction

The objective is to orient the edges of the causal skeleton of the graph in such a way that the likelihood of the data is maximized.

We encode the edges of the net as a string of bits, each one indicating the direction of the relation. For example, the edges of the graph in Figure 1 should be encoded as the string $\{v_1, \dots, v_4\}$, with $v_i \in \{0, 1\}$.

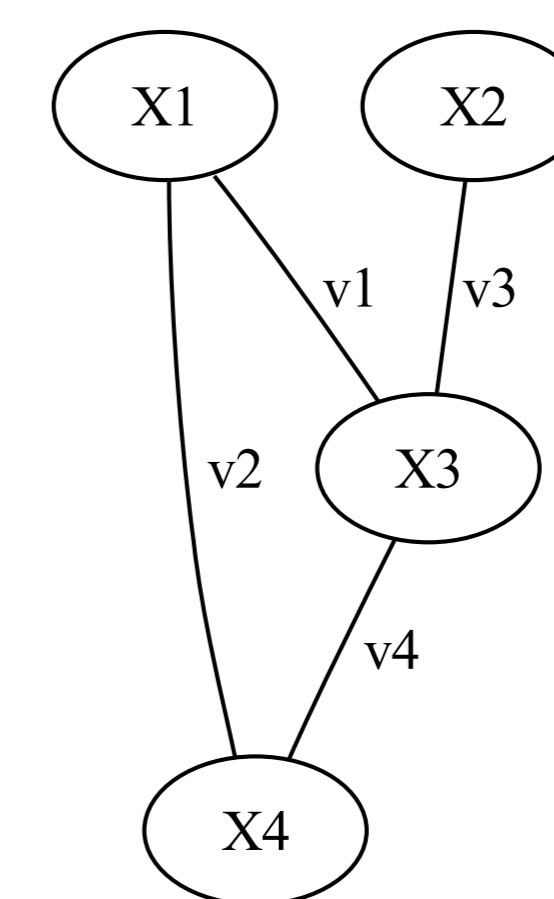


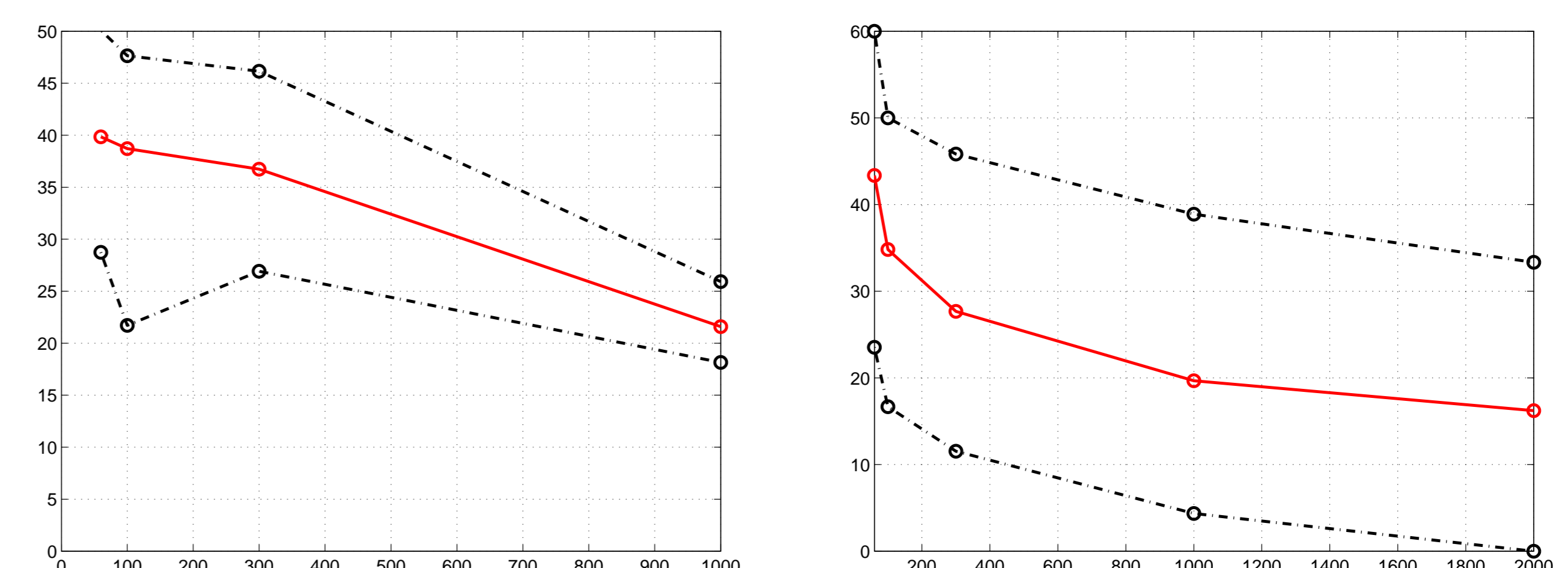
Figure 1: Example of coding with GA

The fitness function is a special case of likelihood, called BDeu, which stands for Bayesian Dirichlet with uniform a priori distribution. The BDeu formula is:

$$P(\mathbf{X}|G) = \prod_{i=1}^n \prod_{j=1}^{q_i} \frac{\Gamma(m'_{ij})}{\Gamma(m'_{ij} + m_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(m'_{ijk} + m_{ijk})}{\Gamma(m'_{ijk})} \quad (1)$$

4. Simulations

The artificial data set is created from Bayesian nets of 15 variables, with binary nodes and Gaussian distribution, and an average number of 20 edges. The conditional probabilities are chosen at random. Figure 2 shows the error probability for both algorithms. The confidence intervals have been figure out using bootstrap resampling. Both algorithms seem to be very sensible to the number of samples.



(a) Orienting edges with classifiers (b) Orienting edges with a GA

Figure 2: Error probability of directions and 90% confidence interval vs sample size.

References

- [1] C.F. Aliferis, I. Tsamardinos, and A. Statnikov. HITON: a novel Markov blanket algorithm for optimal variable selection. In *Proceedings of the AMIA 2003 Annual Symposium*, 2003.
- [2] W.J. Conover. *Practical Nonparametric Statistics*. John Wiley & Sons, 1998.
- [3] J. Pearl. *Causality. Models, Reasoning and Inference*. Cambridge University Press, 2000.
- [4] J.P. Pellet and A. Elisseeff. Using markov blankets for causal structure learning. *Journal of Machine Learning Research*, 9:1295–1342, 7 2008.
- [5] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction and Search*. The MIT Press, 2000.