Causality-Guided Feature Selection

Conference Paper *in* Lecture Notes in Computer Science · November 2016

DOI: 10.1007/978-3-319-49586-6_26

Mandar Chaudhary
North Carolina State University
6 PUBLICATIONS **15** CITATIONS
SEE PROFILE

Doel L. Gonzalez
North Carolina State University
7 PUBLICATIONS **13** CITATIONS
SEE PROFILE

M. P. Angus
North Carolina State University
6 PUBLICATIONS **19** CITATIONS
SEE PROFILE

Dhara Desai
North Carolina State University
2 PUBLICATIONS **4** CITATIONS
SEE PROFILE

Causality-Guided Feature Selection

Mandar S. Chaudhary[1], Doel L. Gonzalez II[1], Gonzalo A. Bello[1], Michael P. Angus[1], Dhara Desai[1], Steve Harenberg[1], P. Murali Doraiswamy[2], Fredrick H. M. Semazzi[1], Vipin Kumar[3], and Nagiza F. Samatova[1,4(✉)], for the Alzheimer's Disease Neuroimaging Initiative*

[1] North Carolina State University, Raleigh, NC, USA
[2] Duke University, Durham, NC, USA
[3] University of Minnesota, Twin Cities, MN, USA
[4] Oak Ridge National Laboratory, Oak Ridge, TN, USA
samatova@csc.ncsu.edu

Abstract. Identifying meaningful features that drive a phenomenon (response) of interest in complex systems of interconnected factors is a challenging problem. Causal discovery methods have been previously applied to estimate bounds on causal strengths of factors on a response or to identify meaningful interactions between factors in complex systems, but these approaches have been used only for inferential purposes. In contrast, we posit that interactions between factors with a potential causal association on a given response could be viable candidates not only for hypothesis generation but also for predictive modeling. In this work, we propose a causality-guided feature selection methodology that identifies factors having a potential cause-effect relationship in complex systems, and selects features by clustering them based on their causal strength with respect to the response. To this end, we estimate statistically significant causal effects on the response of factors taking part in potential causal relationships, while addressing associated technical challenges, such as multicollinearity in the data. We validate the proposed methodology for predicting response in five real-world datasets from the domain of climate science and biology. The selected features show predictive skill and consistent performance across different domains.

1 Introduction

Complex systems, such as the climate and biological systems, are characterized by an intricate interconnected network of interacting factors. These interactions often represent causal relationships among the factors (predictors), as well as between the factors and a phenomenon of interest (response). For example, in

*Data used in preparation of this article were obtained from the Alzheimers disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf the climate science domain, factors may represent large scale ocean and atmospheric patterns, summarized as time series called *climate indices*, whose interactions are known to influence extreme weather phenomena, such as droughts and floods [3,4]. Similarly, in biology, factors may represent genes, the expression levels of which have been found to have an effect on a phenotype of interest, such as disease status [13]. Identifying meaningful factors in these complex systems that can be used to predict the response is a challenging task.

Traditionally, causality-driven methods have been applied to investigate causal relationships between variables. In the climate science domain, they have been used to construct causal graphs (Definition 1) that capture interactions among climate indices [9,10]. The relationships found in these causal graphs are further studied to confirm existing hypotheses and, if possible, generate new ones. On the other hand, in biology, cause-effect relationships are established by performing randomized gene knock-out experiments. Estimating bounds on the potential causal effects of genes on a phenotype has been helpful for prioritizing such experiments [13,14]. However, these approaches require further domain expertise to validate the results and do not focus on identifying predictive factors for any specific response.

Recent methods construct the local causal structure for a response to select predictive features [1,17,20]. These features are identified by utilizing the idea of constraint-based and score-based learning methods. However, these methods do not incorporate the causal effect of a predictor on the response for selecting predictive features and do not identify meaningful cause-effect relationships between variables in the system.

Consequently, we introduce the problem of *causality-guided feature selection* for identifying predictors with significant causal effect on the response. To do this, we construct causal graphs using a constraint-based learning algorithm, such as PC-stable [8], and leverage causal relationships in this graph to estimate the causal effect of a predictor on the response. We evaluate the stability of each predictor by performing a random permutation test to obtain a set of predictors having statistically significant causal effect on the response. In the end, we cluster the predictors and select features from each cluster with the most significant causal effect to form the new feature space.

Finally, we validate our proposed methodology on two motivational use cases in the domains of climate science and biology. Specifically, we apply this methodology to select features for predicting seasonal rainfall in the regions of African Sahel and East Africa, predicting riboflavin production rate in bacterium *B. Subtilis*, and predicting the cognitive score in male and female patients respectively. In climate science, the African Sahel region has been studied extensively following a series of severe droughts in the 1970s and 1980s [4]. Repeated droughts throughout the 2000s have led to a humanitarian crisis in the region, with approximately 10.3 million food-insecure people in 2013[1]. East Africa is a similarly vulnerable region, including within it Lake Victoria, a mostly precipitation-fed

[1] `http://www.fao.org/fileadmin/user_upload/emergencies/docs/SITUATION%20UPDATE%20Sahel%201%2007%202013.pdf`

resource for millions of people. In biology, identifying genes having significant causal effect on a phenotype of interest such as the riboflavin (vitamin B_2) production rate is a challenging task [5,14]. Another important task is to identify biomarkers that can be used to detect the phase of mild cognitive impairment (MCI) which is preceded by Alzheimers disease (AD) among individuals. Currently, no biomarkers have been validated for predicting the risk of AD, and hence there is a greater need to discover key biomarkers [6].

2 Problem Statement

Let $X = \{X_1,\ X_2,\ ..,\ X_p,\ Y\}$ be a set of variables consisting of p predictors, $\{X_1,\ X_2,\ ..,\ X_p\}$, and a response, Y. For example, for our use case in the climate science domain, X may be a set of p climate indices and Y may be seasonal rainfall at a target region (e.g., the African Sahel or East Africa). Informally, we define *causality-guided feature selection* as the task of selecting features based on the *potential causal relationships* among the variables in X, with the goal of improving the prediction of Y. To do this, we first introduce the concepts of *causal graph* and *causal effect*.

Definition 1. (Causal Graph) *Given a set of variables X, a causal graph $G = (V, E)$ is defined as a graph where $V = X$ is the set of nodes and E is the set of edges, such that each directed edge, $X_i \rightarrow X_j$, represents a potential causal relationship where X_i is a potential cause of X_j, and each undirected edge, $X_i - X_j$, or bidirected edge, $X_i \leftrightarrow X_j$, represents an ambiguous relationship between X_i and X_j.*

Definition 2. (Causal Effect) *Given a predictor X_i and a response Y in a causal graph G, the causal effect of X_i on Y is defined as the change in Y for a unit change in X_i. Then, the estimated causal effect of X_i on Y is given by the regression coefficient of X_i, θ_i, when Y is regressed on X_i and its parents S_i; that is,*

$$Y = \theta_i X_i + \theta_{S_i}^{\top} S_i + \epsilon_i \tag{1}$$

where ϵ_i is the residual of Y.

Finally, we formally define the problem of *causality-guided feature selection*: Given a set of variables X consisting of p predictors and a response Y, a causal graph G, and a set of causal effects of the predictors in X on Y, cluster the predictors in X based on their causal effect, and select predictors (i.e., features) from each cluster with the most statistically significant causal effect on the response.

3 Method

In this section, we describe our causality-guided feature selection methodology, as outlined in Algorithm 1. First, we use a constraint-based learning algorithm to

Algorithm 1 Causality-Guided Feature Selection

```
Require: A set of variables $X = \{X_1, X_2, .., X_p, X_{p+1}\}$ consisting of $p$ predictors,
    $\{X_1, X_2, ..., X_p\}$, and a response, $Y = X_{p+1}$
 1: Let $f_{new} = \emptyset$ be the new feature space
 2: Let $G$ be a CPDAG constructed using the PC-stable algorithm (see Section 3.1)
 3: Let $C$ be the set of potential causal relations in $G$
 4: Let $\mathit{\Theta} = \emptyset$ be a set of statistically significant causal effects
 5: Let $\mathit{\Phi}$ be a set of p-values of the statistically significant causal effects
 6: for each $c = X_i \to X_j$ in $C$ do
 7:     Let $\mathit{\Theta}_i = \emptyset$
 8:     Let $\mathcal{G}$ be the set of Markov equivalent graphs generated for $X_i$
 9:     for each $g \in \mathcal{G}$ do
10:         Let $\theta_{X_{i,1}}$ be the causal effect of $X_i$ on $Y$ computed with PCR using $X_i$ and
            its set of parents $S_i$ in $g$ (see Section 3.2)
11:         $[\theta_{X_{i,1}}$, $\textit{p-value}_{\theta_{i,1}}]$=ASSESS_STABILITY($\theta_{X_{i,1}}$, $X_i$, $S_i$, $Y$)
12:         $\mathit{\Theta}_i = \mathit{\Theta}_i \cup \theta_{X_{i,1}}$
13:     end for
14:     if $\mathit{\Theta}_i \neq \emptyset$ then
15:         $\theta_i = \arg\min_{\theta \in \mathit{\Theta}_i} |\theta|$
16:         $\mathit{\Theta} = \mathit{\Theta} \cup \theta_i$
17:         $\mathit{\Phi} = \mathit{\Phi} \cup$ $\textit{p-value}_{\theta_i}$
18:     end if
19:     Repeat steps 7-18 for $X_j$ and $\mathit{\Theta}_j$ if $X_j \neq Y$
20: end for
21: $f_{new}$ =FEATURE_SELECTION($\mathit{\Theta}$, $\mathit{\Phi}$, $X$)
22: return $f_{new}$
```

Algorithm 2 ASSESS_STABILITY

```
Require: A causal effect, $\theta_{X_{i,1}}$, a predictor, $X_i$, its parents $S_i$, and a response $Y$
Let $N$=100
Let $\mathit{\Theta}_{rand}$ be a set of randomized causal effects computed from $N$ random permu-
tations of the response (see Section 3.2)
$\textit{p-value} = \frac{\left|\{\theta'_m \in \mathit{\Theta}_{rand} \text{ s.t. } |\theta'_m| \geq |\theta_{X_{i,1}}|\}\right| + 1}{N+1}$
if $p$-value $< 0.05$ then
    return $[\theta_{X_{i,1}}$, $\textit{p-value}]$
else
    return $\emptyset$
end if
```

construct a causal graph and select the potential causal relationships in the graph (Section 3.1). Second, for each predictor that takes part in a potential causal relationship, we compute its causal effect on the response using a methodology that addresses multicollinearity and then estimate the significance of this causal effect by performing a random permutation test (Section 3.2). Finally, we cluster the predictors based on their causal effect and select features from each cluster with the most statistically significant causal effect on the response (Section 3.3).

Algorithm 3 FEATURE_SELECTION

```
Require: A set of statistically significant causal effects, $\Theta$, its corresponding set of p-values $\Phi$, and set of predictors in $X$.
1: Perform K-Means clustering on $\Theta$ by identifying the optimal number of clusters, $k$, using the Elbow Method
2: for each cluster $c_i$ do
3:    Let $\phi \subset \Phi$ be the p-values of statistically significant causal effects in $c_i$
4:    Select all predictors, $p_i \subset X$, in $c_i$ whose causal effect has p-value equal to $min(\phi)$
5:    $f_{new} = f_{new} \cup p_i$
6: end for
7: return $f_{new}$
```

3.1 Constructing Causal Graphs and Selecting Potential Causal Relationships

We construct a graph of ambiguous and potential causal relationships using a constraint-based structure learning algorithm. Specifically, we use the PC-stable algorithm because of its ability to construct graphs with order-independent adjacency structure and to mitigate the effect of false positives edges [8]. In the next two paragraphs we present a summary of the constraint-based structure learning algorithm.

In the first step, the PC-stable algorithm constructs a completed undirected graph G over a set of variables X and initializes the size of the conditioning set, m, to zero. Next, for each variable $X_i \in X$, it stores the nodes adjacent to X_i in its adjacency set $a(X_i)$. For every pair of adjacent variables X_i and X_j in G, it checks whether the two variables are independent conditioned on S (i.e., $X_i \perp X_j|S$), such that $S \subseteq a(X_i)$ or $S \subseteq a(X_j)$ and $|S| = m$. If the variables are conditionally independent, the edge between them is removed from G and the conditioning variable(s), S, is stored in their separating set, $sepset(X_i, X_j)$ and $sepset(X_j, X_i)$. Similarly, the remaining pairs of adjacent variables are checked for conditional independence. This completes the first iteration of the conditional independence test. The adjacency set is updated for every variable, and the value of the conditioning set, m, is incremented by 1 in the next iteration. At the end of this step, the algorithm yields a *skeleton* of the causal graph, which contains undirected edges between variables that were not found to be conditionally independent. In our experiments, we use the Fisher's Z test to determine if two variables are conditionally independent at a significance threshold $\alpha = 0.05$.

In the next step, for every unshielded triple $(X_i - X_j - X_k)$ such that X_i and X_k are not adjacent, the algorithm orients $X_i - X_j - X_k$ into a v-structure, $X_i \rightarrow X_j \leftarrow X_k$, if and only if $X_j \notin sepset(X_i, X_k)$. Then the algorithm tries to orient as many remaining edges as possible using the following set of rules:

- Rule 1: Given $X_i \rightarrow X_j$ and $X_j - X_k$, orient $X_j - X_k$ to $X_j \rightarrow X_k$ such that X_i and X_k are not adjacent.
- Rule 2: Given $X_i - X_k$ and a chain $X_i \rightarrow X_j \rightarrow X_k$, orient $X_i - X_k$ into $X_i \rightarrow X_k$.

- Rule 3: Given two chains, $X_i - X_j \rightarrow X_k$ and $X_i - X_l \rightarrow X_k$, orient $X_i - X_k$ into $X_i \rightarrow X_k$.

The output of the PC-stable algorithm is a completed partially directed acyclic graph (CPDAG), which represents an approximation of the Markov equivalence class of the data. An estimated CPDAG can contain directed, undirected, and bidirected edges, as shown in Figure 1a. A directed edge $X_1 \rightarrow Y$ represents a cause-effect relationship where X_1 is a potential cause of Y, an undirected edge $X_1 - X_2$ implies some association between the variables, and a bidirected edge $X_2 \leftrightarrow X_3$ represents a sampling error or a hidden common cause of X_2 and X_3 that is not present in the data.

Assumptions In order to interpret a CPDAG causally, we consider the following three assumptions:

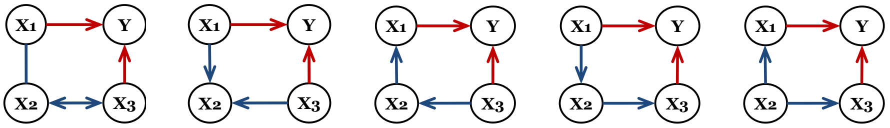
- The underlying causal structure is sparse and acyclic.
- The graph built by the PC-stable algorithm is a well approximated representation of the underlying probability distribution in the data (i.e., causal faithfulness).
- There is absence of confounding variables in the data (i.e., causal sufficiency); i.e., given any two variables X and Y in the data having a common cause Z, then Z is also present in the data.

In real-world scenarios, it is difficult to assume that a system is causally sufficient, since there might be factors that are difficult to observe and measure. As a result, we cannot prove the existence of the potential causal relationships in the CPDAG. Nonetheless, they can provide an approximation of the underlying interactions between the variables. There are causal inference algorithms such as the Fast Causal Inference (FCI) algorithm and the Really Fast Causal Inference (RFCI) that detect the presence of hidden causes [19].For future work, we plan to explore and determine the applicability of such causal inference algorithms to our proposed methodology.

Once the CPDAG is constructed using the PC-stable algorithm, we select all the directed edges in the graph for further processing because they represent potential causal relationships between pairs of variables. Moreover, these directed edges are present across all the Markov equivalent DAGs that can be generated from the estimated CPDAG. Figure 1a shows a CPDAG and a set of Markov equivalent DAGs (Figures 1c to 1e) generated from the CPDAG by orienting all the undirected and bidirected edges into directed edges such that no additional v-structure or cycle is created.

3.2 Estimating Causal Effects and Assessing its Statistical Significance

We estimate the causal effect on the response of every predictor that takes part in a potential causal relationship. For example, consider a potential causal relationship $X_i \rightarrow X_j$, where an external intervention on predictor X_i leads to a unit change in X_i. To estimate the causal effect of X_i on Y, the Intervention



(a) CPDAG (b) D1 (c) D2 (d) D3 (e) D4

Fig. 1: (a) A completed partially directed acyclic graph (CPDAG) and (b)-(e) the set of possible DAGs D1-D4. D1 is an invalid DAG as it contains an additional v-structure, $X_1 \rightarrow X_2 \leftarrow X_3$, that is not present in the estimated CPDAG. D2-D4 are valid DAGs and they belong to the same Markov equivalence class.

calculus when DAG is Absent (IDA) method has been proposed, which explores the local neighborhood of X_i by changing every undirected and bidirected edge incident on X_i to a directed edge [14]. Thus, if there is a total of m undirected and bidirected edges incident on X_i, then up to 2^m Markov equivalent graphs can be generated. For each graph, the IDA method identifies the parents of X_i (i.e., predictors having a directed edge towards X_i). The response Y is then regressed on X_i and its parents, $S_i \subseteq \{X_1, X_2, ..., X_p, Y\} \setminus X_i$; that is,

$$Y = \theta_i X_i + \theta_{S_i}^{\top} S_i + \epsilon_i \tag{2}$$

where the regression coefficient of X_i, θ_i, is the estimated causal effect of X_i on Y for the corresponding Markov equivalent graph and ϵ_i is the residual of Y.

Thus, the IDA method yields a set of causal effects for X_i, $\Theta_i = \{\theta_i^1,\ \theta_i^2\ , ...,\ \theta_i^k\}$, across $k \leq 2^m$ Markov equivalent DAGs. The causal effect θ_i is 0 if $Y \in S_i$, since no change in X_i can have an effect on its parents; otherwise, θ_i is estimated as shown above.

For several real-world applications, such as our use cases in the domains of climate science and biology, the data exhibits a high degree of multicollinearity; that is, near-linear dependencies among the predictors. We determine the presence of multicollinearity by computing the variance inflation factor (VIF); for example, for our climate data set, a high VIF (i.e., greater than 10) indicates a high degree of multicollinearity in the data [16]. The presence of multicollinearity leads to a poor estimation of the coefficients in the linear regression models in Equation 2 [15]. To address this issue, we perform Principal Component Regression (PCR), instead of linear regression, to estimate the causal effects.

Principal Component Regression Given a Markov equivalent graph, let $\mathcal{X}_i$ be a matrix whose columns consist of predictor X_i and its parents (i.e., the predictors in S_i). Then, Principal Component Regression (PCR) computes n principal components of $\mathcal{X}_i$, where n is less than or equal to the number of columns of $\mathcal{X}_i$, using singular value decomposition (SVD) as follows:

$$\mathcal{X}_i = T_i {P_i}^{\top} + \epsilon_{\mathcal{X}_i} \tag{3}$$

where T_i is the score matrix, P_i is the loading matrix and $\epsilon_{\mathcal{X}_i}$ is the unexplained variance of $\mathcal{X}_i$. Next, PCR builds a regression model with Y as the response and the score matrix T_i as the predictors; that is,

$$Y = \theta_{T_i} T_i + \epsilon_{T_i} \tag{4}$$

We solve for the regression coefficients θ_{T_i} of the score matrix T_i using the least squares method as follows:

$$\theta_{T_i} = (T_i^{\top} T_i)^{-1} T_i^{\top} Y \tag{5}$$

The regression coefficients of the score matrix T_i, θ_{T_i}, are then used to compute the regression coefficients of the original predictor matrix $\mathcal{X}_i$ as follows:

$$\theta_{\mathcal{X}_i} = P_i \theta_{T_i} \tag{6}$$

where $\theta_{\mathcal{X}_i}$ contains the regression coefficients of X_i and of the predictors in S_i across n principal components. We then select, as the causal effect of X_i on Y, the regression coefficient of X_i that corresponds to the first principal component, $\theta_{X_{i,1}}$, since it captures the maximum variance of $\mathcal{X}_i$.

Statistical Significance Test To assess the significance of a causal effect $\theta_{X_{i,1}}$, we perform the statistical test described in Algorithm 2. The response Y is randomly permuted $N = 100$ times and the corresponding set of randomized causal effects $\Theta_{rand} = \{\theta_1^{'}, \theta_2^{'}, ..., \theta_N^{'}\}$ is computed using Equation 6. A p-value is calculated to measure the probability that the magnitude of a randomized causal effect is greater than or equal to the magnitude of $\theta_{X_{i,1}}$. The causal effect $\theta_{X_{i,1}}$ is considered to be statistically significant if its p-value is less than 0.05.

This test is performed for each causal effect of X_i on Y computed across all the Markov equivalent graphs. The result is a set of statistically significant causal effects of X_i on Y, $\Theta_i = \{\theta^1_{X_{i,1}}, \theta^2_{X_{i,1}}, ..., \theta^l_{X_{i,1}}\}$, where $l \leq k$. Similarly, we compute the set of statistically significant causal effects of X_j on Y, Θ_j, for predictor X_j taking part in the potential causal relationship $X_i \rightarrow X_j$. From a set of statistically significant causal effects, Θ_i, we select θ_i, such that

$$\theta_i = \arg\min_{\theta \in \Theta_i} |\theta| \tag{7}$$

The causal effect θ_i represents the minimum change in the response Y for a unit change in X_i. We compute the causal effects θ_i for every predictor X_i that takes part in a potential causal relationship in the CPDAG. This results in a set of predictors having a statistically significant causal effect on the response.

3.3 Feature Selection via Clustering

Finally, as a feature selection technique, we group the predictors by clustering them based on their causal effects, as described in Algorithm 3. Specifically, we use the K-Means clustering method and identify the optimum number of clusters k using the Elbow Method. From each cluster, we select the predictors with the most statistically significant causal effect on the response i.e., the ones with the lowest p-value from the statistical test (see Section 3.2). By doing so, we minimize the number of redundant predictors having similar causal effect on the response.

4 Empirical Evaluation

In this section, we describe the application of our proposed causality-guided feature selection methodology. Specifically, our goal is to select features that improve the prediction of 1) seasonal rainfall at the African Sahel and East Africa regions, 2) riboflavin production rate, and 3) cognitive score of male and female patients.

4.1 Data Description

We used two real-world data sets from the climate science domain and three real-world data sets from the biology domain to demonstrate the performance of our methodology.
Climate Science: For the African Sahel region (ASR), we used seasonal rainfall from July to September as the response and monthly values from January to June of 34 climate indices as the predictors. Similarly, for the region of East Africa (EAR), we used seasonal rainfall from October to December as the response and monthly values from January to September of 33 climate indices as the predictors. Monthly rainfall data for both regions was obtained from the Gridded Precipitation Climatology Centre (GPCC) V6 data set [18]. Monthly data for 30 climate indices was obtained from the Earth System Research Laboratory (ESRL)[1], and 5 additional climate indices were constructed to incorporate local atmospheric conditions at each region. For our experiments, we used 57 years of data from 1951 to 2007.
Biology: For our first data set, we used publicly available data[2] containing 71 observations of 101 variables measuring the logarithm of the expression level of 100 genes and the riboflavin production rate (RPR) in the bacterium *B. subtilis* [5]. Our second and third data sets were collected from the Alzheimer's Disease Neuroimaging Initiative (ADNI). The data sets contained microarray gene expression data collected from 266 male patients and 219 female patients. The cognitive score of the subjects in each gender was used as the response variable, i.e., CS_M for male patients and CS_F for female patients.

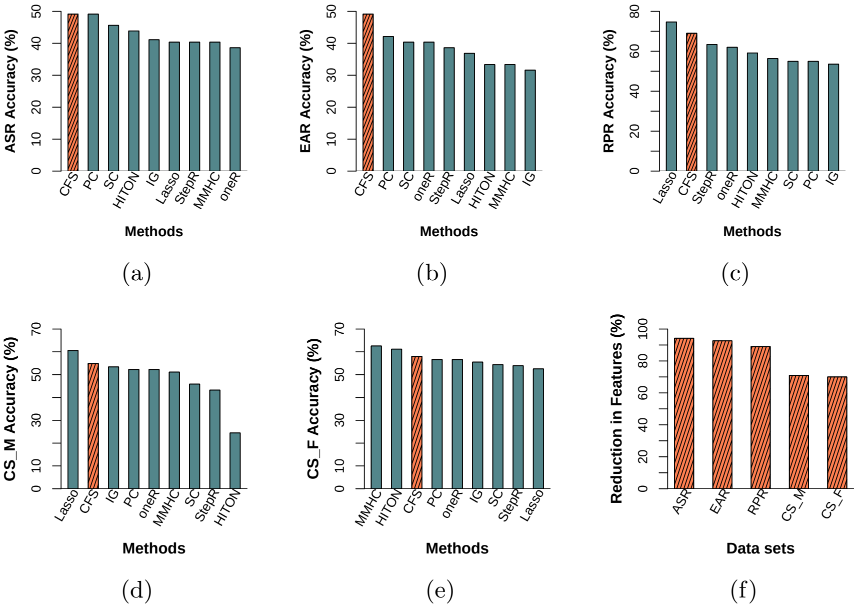
4.2 Data Preprocessing

The following pre-processing steps were performed after dividing the data into training and test sets using leave-one-out cross-validation (LOOCV),

1. **De-trending**: Given the temporal nature of the climate data sets, the predictors and the response variable were detrended to remove seasonal trends. This step was performed only on climate data sets.
2. **Normalization**: The predictors and the response variable were standardized using their z-scores. Note that the z-scores were computed using the average and the standard deviation from the training sets.

[1] `http://www.esrl.noaa.gov/psd/data/climateindices/list/`

[2] `http://www.annualreviews.org/doi/suppl/10.1146/annurev-statistics-022513-115545/suppl_file/riboflavinv100.csv`



(a) (b) (c)

(d) (e) (f)

Fig. 2: Mean classification accuracy over leave-one-out cross validation (Accuracy) for the African Sahel (2a East Africa (2b), *B. subtilis* (2c), diagnostic information for Male (2d) and Female (2e) obtained from Causality-Guided Feature Selection (CFS), Max-Min Hill Climbing (MMHC), HITON Markov Blanket (HITON) Spearman Correlation (SCorr), Pearson Correlation (PCorr), oneR (oneR), Lasso Regression (Lasso), Stepwise Regression (StepR), and Information Gain (IG). 2f shows the percentage reduction in the features using CFS across five data sets.

Due to the indefinite amount of time taken by the PC-stable algorithm for constructing causal graphs from the microarray gene expression data, we used the training set to select the top-100 genes correlated with the response as predictors.

4.3 Performance Comparison

We evaluate the performance of our causality-guided feature selection methodology by training classification models using C5.0 decision trees and regression models using the linear regression method. The response variables for ASR, EAR and RPR data sets were discretized into three categories: high, normal or low (i.e., values in the higher 66.7^{th} percentile, between the lower 33.3^{rd} and the higher 66.7^{th} percentile, and in the lower 33.3^{rd} percentile, respectively). For the two ADNI data sets, CS_M and CS_F, the cognitive score collected from the Mini-Mental State Exam (MMSE) was used as the continuous response variable. Additionally, there were two groups of patients within each gender based on the diagnostic information made available by ADNI. The first group comprised

Table 1: RMSE scores for prediction of seasonal rainfall at African Sahel (ASR) and East Africa (EAR), riboflavin production rate (RPR) and cognitive score for male (CS_M) and female (CS_F) patients obtained from all the feature selection methods. RMSE scores within 6% of the best performing method are highlighted in bold. (*) indicates that a feature selection method could not find any feature from some of the training sets during cross-validation

Data set	Feature Selection Methods								
	CFS	PCorr	SCorr	oneR	StepR	IG	Lasso	MMHC	HITON
ASR	**0.9787**	1.1499	1.0942	**0.9985**	1.7362	**0.9978**	1.2304	*	*
EAR	0.9860	1.0720	1.0448	1.3769	1.3885	1.3751	3.5515	**0.7599**	**0.7655**
RPR	**0.5483**	**0.6112**	**0.5335**	0.8288	1.0956	0.6077	0.6432	0.7978	0.7114
CS_M	**1.0576**	1.1976	1.1805	1.1939	1.134	1.2086	1.2039	**1.0477**	**1.0148**
CS_F	**1.1079**	**1.0646**	1.1285	1.1337	1.1859	1.1321	1.2397	**1.1033**	**1.0996**

of controls (i.e., patients who did not suffer from Alzheimer's), and the second group comprised of patients suffering from late Mild Cognitive Impairment (LMCI). This diagnostic information for males and females was used as the discretized response variable in both the data sets. The decision trees were trained on these five data sets using the corresponding discretized response variables. The regression models were built using the z-scores of the response variables.

The performance of our proposed methodology was compared against eight feature selection methods, including univariate methods, two regression methods, and two local causal discovery-based methods. For the univariate methods, we selected the top-K predictors where the value of K was equal to the average number of features constructed by our proposed methodology across all the folds.

The classification accuracies of the feature selection methods on five real-world data sets are shown in Figure 2a-2e. We observe that our proposed methodology is the best performing method for ASR and EAR, and its accuracy is within 6% of the best performing method for RPR, CS_M and CS_F. While Lasso outperforms all the methods on RPR and CS_M, it is the worst performing method for CS_F and our methodology has a minimum improvement of 8% over Lasso for ASR and EAR. Similarly, MMHC and HITON have higher accuracy values while predicting CS_F, but their performance is as good as random guessing for EAR, and HITON is the worst performing method for CS_M.

We now compare the root mean squared errors (RMSE) for each method to evaluate the performance of the selected features on the regression models. Tables 1 presents our findings that four out of five times the RMSE score of our proposed methodology is within 6% of the best performing method. The remaining methods, however, don't show the same level of robustness in terms of the results obtained. For example, Pearson Correlation was among the top performing methods in terms of classification accuracy for ASR and EAR, but its performance is relatively poor when evaluating regression models. Furthermore, although Lasso has higher accuracy values for RPR and CS_M (see Figure 2c-2d), it is the worst performing method for EAR and CS_F. While MMHC and HITON have high accuracy values for CS_F, they had the worst RMSE scores for ASR as they could not find any feature from some of the training sets during cross-validation.

A summary of the results is shown in Table 2. We observe that the classification accuracy and RMSE score of our methodology is consistently within 6% of the best performing method. Note that while methods such as Pearson Correlation, Lasso, MMHC and HITON perform well on a few data sets, they are also the worst performing methods or within 6% of the worst performing method on other data sets. This indicates that their prediction performance varies significantly with data sets from different domains, whereas our methodology shows predictive skill across all of the data sets in terms of classification accuracy and RMSE score. Moreover, we also observe a percentage reduction in the selected features by at least 70% as compared to the original feature space across all the data sets (see Figure 2f).

4.4 Time Complexity

The majority of the time taken by the proposed methodology is due to constructing a causal graph, computing causal effects and their significance. The asymptotic time complexity of the PC-stable algorithm is polynomial time on

Table 2: A summary of the performance of all the feature selection methods in terms of classification accuracy and RMSE scores across all five data sets. A solid upper triangle indicates the best performing method, whereas a solid lower triangle indicates the worst performing method on a given data set. An upper triangle with pattern indicates that a method's performance was within 6% of the accuracy and RMSE score of the best performing method, and a lower triangle with pattern indicates a method's performance was within 6% of the accuracy and RMSE score of the worst performing method.

Performance Metrics	Data set	Feature Selection Methods								
		CFS	PCorr	SCorr	oneR	IG	StepR	Lasso	MMHC	HITON
Accuracy	ASR	◤	◤	◸	◢		◿		◿	◸
	EAR	◤				◢			◿	◿
	RPR	◸	◿	◿		◢		◤		
	CS_M	◸		◿			◿	◤		◢
	CS_F	◸		◿			◿	◢	◤	◸
RMSE	ASR	◤			◸	◸			◢	◢
	EAR							◢	◤	◸
	RPR	◸	◸	◤			◢			
	CS_M	◸	◿	◿	◿	◿		◿	◸	◤
	CS_F	◸	◤	◸			◿	◢	◸	◸

sparse graphs, i.e., $O(p^q)$ where p is the number of variables in the graph and q is the maximum number of vertices adjacent to any vertex in the graph such that $q = O(n^{1-b})$, where n is the sample size and $0 < b \leq 1$. To estimate a causal effect, the time complexity of linear regression is dominated by matrix multiplication; i.e., $O(p^2 n)$ if $n > p$ else it is $O(p^3)$. In the worst case, we may end up computing causal effect of every predictor on the response variable. As a result, when $n > p$ the time taken to compute the causal effect of each predictor on the response variable and its significance across all the Markov equivalent graphs is $O(2p \cdot 2^q \cdot (p^2 n + 100p^2 n)) \approx O(2^{q+1} p \cdot cp^2 n)$ where 2^q is the number of Markov equivalent graphs and c is a constant. Thus, the total time complexity is $O(p^q + 2^{q+1} p \cdot cp^2 n)$. Similarly, for $n < p$ it is $O(p^q + 2^{q+1} p \cdot cp^3)$.

5 Related Work

The underlying behavior of a complex system can be attributed to the intricate network of potential cause-effect relationships between factors. Identifying meaningful predictors that can further the understanding of such complex systems is a challenging task. In the climate science domain, constraint-based structure learning methods for causal discovery have been applied to generate graphs of information flow (i.e., causal graphs) describing the interactions between four climate indices [9,10]. Similarly, in the domain of bioinformatics, causal inference algorithms have been applied to estimate the brain network structure from fMRI data and to explain the variations observed in high-throughput gene expression data [7,12]. Note that these methods focus on studying potential causal relationship using domain knowledge, which may not be feasible for high-dimensional systems.

Feature selection methods based on local causal structure learning identify variables within the Markov Blanket of a target variable [1,17,20,11]. HITON and Max-Min Hill Climbing (MMHC) are prominent examples of such feature selection methods, which employ the divide and conquer approach to find the parents, children and the spouses of the target variable [1,2]. However, these methods are based on constraint-based learning and do not incorporate the causal information between the variables to select the features. In this work, we have proposed a novel method that integrates the causal strength of the predictors on the target variable to select meaningful predictors. The results show that our method produces better predictive performance over these methods.

6 Conclusion

Causality-guided methods have been used in multiple domains to facilitate the understanding of complex systems. Traditionally, the application of these methods has been limited to descriptive and inferential purposes. In this work, we propose a causality-guided feature selection methodology that identifies predictors taking part in potential causal relationships and selects features based on their causal strength with respect to the response via clustering. We achieve this through a number of technical contributions, such as estimating the statistical significance of causal effects on the response, while addressing multicollinearity in the data. Our proposed methodology was found to perform consistently in terms of classification accuracy and RMSE score across real-world data sets from the domains of climate science and biology, suggesting that the newly selected features have predictive skill for the response.

Acknowledgments This material is based upon work supported in part by the Laboratory for Analytic Sciences (LAS), the Department of Energy National Nuclear Security Administration under Award Number(s) DE-NA0002576 and NSF grant 1029711. Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie,Alzheimers Association; Alzheimers Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimers Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California. PMD has received research grants and/or advisory fees from several government agencies, advocacy groups and pharmaceutical/imaging companies, and received a grant from ADNI to support data collection for this study. He also owns stock in several companies whose products are not discussed here. The final publication is available at Springer via http://dx.doi.org/10.1007/978-3-319-49586-6_26

References

1. Aliferis, C.F., Statnikov, A., Tsamardinos, I., Mani, S., Koutsoukos, X.D.: Local causal and markov blanket induction for causal discovery and feature selection for classification Part I: Algorithms and empirical evaluation. J. Mach. Learn. Res. 11, 171–234 (2010)

2. Aliferis, C.F., Tsamardinos, I., Statnikov, A.: Hiton: a novel markov blanket algorithm for optimal variable selection. In: AMIA Annu. Symp. Proc. vol. 2003, p. 21 (2003)
3. Andrews, E., Antweiler, R.C., Neiman, P.J., Ralph, F.M.: Influence of enso on flood frequency along the california coast. J. Climate 17(2), 337–348 (2004)
4. Bader, J., Latif, M.: The 1983 drought in the west sahel: a case study. Climate Dynam. 36(3), 463–472 (2011)
5. Bühlmann, P., Kalisch, M., Meier, L.: High-dimensional statistics with a view toward applications in biology. Annu. Rev. of Stat. Appl. 1, 255–278 (2014)
6. Chen, Z., Padmanabhan, K., Rocha, A.M., Shpanskaya, Y., Mihelcic, J.R., Scott, K., Samatova, N.F.: Spice: discovery of phenotype-determining component interplays. BMC Syst. Biol. 6(1), 1–19 (2012)
7. Chindelevitch, L., Ziemek, D., Enayetallah, A., Randhawa, R., Sidders, B., Brockel, C., Huang, E.S.: Causal reasoning on biological networks: interpreting transcriptional changes. Bioinformatics 28(8), 1114–1121 (2012)
8. Colombo, D., Maathuis, M.H.: Order-independent constraint-based causal structure learning. J. Mach. Learn. Res. 15(1), 3741–3782 (2014)
9. Ebert-Uphoff, I., Deng, Y.: Causal discovery for climate research using graphical models. J. Climate 25(17), 5648–5665 (2012)
10. Ebert-Uphoff, I., Deng, Y.: Causal discovery from spatio-temporal data with applications to climate science. In: Proc. of the 2014 13th Int Conf. on Machine Learning and Applications. pp. 606–613. IEEE (2014)
11. Guyon, I., Aliferis, C., Elisseeff, A.: Causal feature selection. Computational methods of feature selection pp. 63–86 (2007)
12. Iyer, S.P., Shafran, I., Grayson, D., Gates, K., Nigg, J.T., Fair, D.A.: Inferring functional connectivity in mri using bayesian network structure learning with a modified pc algorithm. Neuroimage 75, 165–175 (2013)
13. Maathuis, M.H., Colombo, D., Kalisch, M., Bühlmann, P.: Predicting causal effects in large-scale systems from observational data. Nature Methods 7(4), 247–248 (2010)
14. Maathuis, M.H., Kalisch, M., Bühlmann, P., et al.: Estimating high-dimensional intervention effects from observational data. Ann. of Stat. 37(6A), 3133–3164 (2009)
15. Montgomery, D.C., Peck, E.A., Vining, G.G.: Introduction to linear regression analysis. John Wiley & Sons (2015)
16. Neter, J., Kutner, M.H., Nachtsheim, C.J., Wasserman, W.: Applied linear statistical models, vol. 4. Irwin Chicago (1996)
17. Peña, J.M., Nilsson, R., Björkegren, J., Tegnér, J.: Towards scalable and data efficient learning of markov boundaries. Int. J. Approx. Reason. 45(2), 211–232 (2007)
18. Schneider, U., Becker, A., Finger, P., Meyer-Christoffer, A., Rudolf, B., Ziese, M.: Gpcc full data reanalysis version 6.0 at 0.5: monthly land-surface precipitation from rain-gauges built on gts-based and historic data. FD_M_V6_050 (2011)
19. Spirtes, P., Glymour, C.N., Scheines, R.: Causation, prediction, and search, vol. 81. MIT press (2000)
20. Tsamardinos, I., Aliferis, C.F., Statnikov, A.: Time and sample efficient discovery of markov blankets and direct causal relations. In: Proc. of the Ninth ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining. pp. 673–678. ACM (2003)