Introduction to Artificial Intelligence 10 (2019) Submitted 2019/12/26; Published 2019/12/26

A Survey towards Causality in Machine Learning

Feng Shi FENGS19@MAILS.TSINGHUA.EDU.CN
Class:91
Institute for Interdisciplinary Information Sciences
Tsinghua University, Beijing, China

Abstract

Causality is an important concept in machine learning. However, it has not been attached enough importance. However, in the future, artificial intelligence can not be merely curve fitting and stay at the level of association. Causality models will be a significant step to achieve strong AI. In this survey, the definition of causality will be introduced at first, including three levels of causality and the difference between event causality and procedure causality. In the third section, a model called SCM which is the most complete causal model at present will be introduced. Then we will demonstrate how can it solve some of the problems in machine learning. A simple approach to the causal loop diagram will be made in the fifth part. Although researches towards causality is not focused and many parts need to be improved, many remain optimistic about the development of causality.

Keywords: Machine Learning, Causality, SCM, Bayesian Network, CLD

1. Introduction

Causality, which is a concept that troubles philosophers for hundreds of years, have considered being an important approach to a breakthrough of machine learning. According to Judea Pearl who proposed the Bayesian Network, causality can change the basic idea of conventional machine learning methods.

Nowadays, most of the machine learning models are based on statistic models. They are actually accurate curve fitting of data and can not render the machines abilities of thinking or acting like animals. For example, in the area of image recognition, machines already make a great success. However, the success is based on mass data (millions of manually annotated pictures), high computing consumption (hundreds of GPUs) and systems with large storage. The most important shortage is that problems like image recognition are IID (Independent and Identically Distributed) or manually made to IID (Schölkopf, 2019). This shortage leads to the problem of adversarial vulnerability, which means if we make some minor changes to the property of IID, the accuracy of the models will sharply decrease. For instance, most of the image recognition systems are probably hard to identify a fish in the grass.

These problems can be solved by the models based on causality for the causality is able to contain much more information. The models will not stay at the level of association, the fish is not identified just because most of the similar pictures contain a fish. Causality models give the machine the ability to reason and solve piles of problems that seems to be unsolvable or extremely hard for conventional machine learning models. Actually, causality

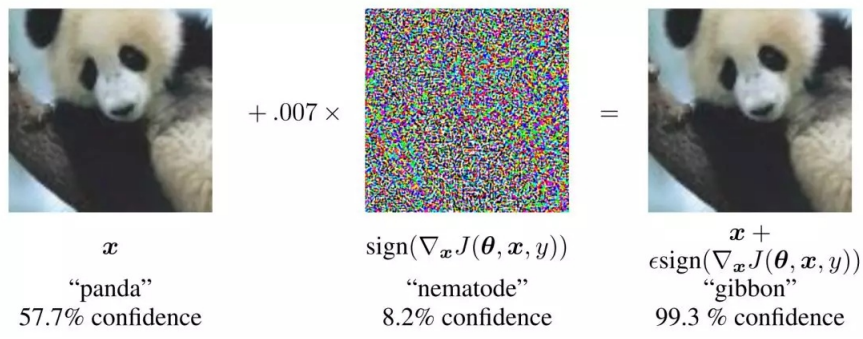

Figure 1: Adversarial Vulnerability Makes Machine Hard to Identify the Panda (Ilyas et al., 2019).

can be a significant approach to the concept of strong AI. "I expect this symbiosis to yield systems that communicate with users in their native language of cause and effect and, leveraging this capability, to become the dominant paradigm of next-generation AI" (Pearl, 2019). In this survey, we will have a summary of the history of causality at first and then introduce the SCM model brought out by Judea Pearl. We will have a summary of the improvements due to the SCM model and a simple introduction of the CLD model. In the last part of the survey, there will be a vision of future developments and some big open problems.

2. The Definition of Causality

To begin with, we point out causality is hard to define. We will see that most of the conventional definitions are not completely right. In order to define causality, we need to find a method to determine causal.

2.1 Conventional Definition of Causality

At first, we will demonstrate the conventional definitions of causality. Scottish Enlightenment philosopher David Hume defined causality as "constant conjunction"(Dicker, 2002). It means that if event A always happens before B and A and B are relevant, then A is the reason of B. It is obviously wrong. For example, squirrels always store food before the coming of winter. However, if we force all squirrels in the world not to store food, winter will still come anyhow. Therefore, the behavior of squirrels is not the reason why we have a season with low temperatures and heavy snow.

Another attempt is by Karl Pearson. Pearson came up with the concept of statistical correlation (Dempster, 1990). However, it can not be used as a definition of causality. The definition of correlation is as below (Mari and Kotz, 2001):

$$corr(A, B) = \frac{cov(A, B)}{\sigma_A \sigma_B} = \frac{E[(A - \mu_A)(B - \mu_B)]}{\sigma_A \sigma_B} \tag{1}$$

A,B are events. cov is the covariance. σ_A and σ_B are the standard deviations of A and B, correspondingly. μ_A and μ_B are means of A and B, correspondingly. We see from it that the function $corr$ is symmetric for A and B, which is contradictory to the property of causality. We can have a typical example to illustrate this. The sales volume of ice-cream is positively related to the number of forest fires in California but it not because there is a causal relation between them, the reason is only that the summer is coming as is shown in figure 2. This leads to the result that statistic model cannot completely express the causality. Actually, it can only lead to the INUS condition (an insufficient but necessary part of a condition which is itself unnecessary but sufficient for the result).

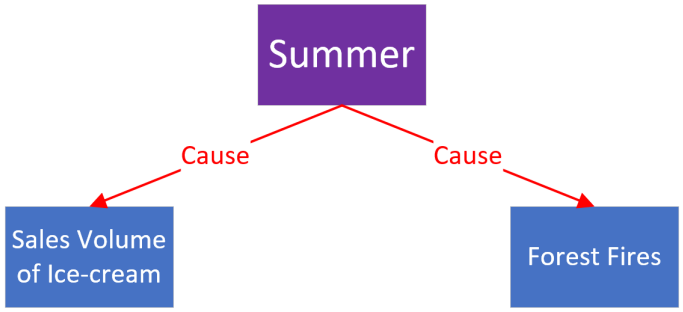

Figure 2: The Relation between Sales Volume of Ice-cream and Forest Fires in California.

2.2 Three Levels of Causality Models

Actually, there are three levels of causality according to Judea Pearl. They're association, intervention, and counterfactuals (Pearl, 2019) as is shown in figure 3.

Level (Symbol)	Typical Activity	Typical Questions	Examples
1. Association $P(y\|x)$	Seeing	What is? How would seeing X change my belief in$Y$?	What does a symptom tell me about a disease? What does a survey tell us about the election results?
2. Intervention $P(y\|do(x), z)$	Doing Intervening	What if? What if I do X?	What if I take aspirin, will my headache be cured? What if we ban cigarettes?
3. Counterfactuals $P(y_x\|x', y')$	Imagining, Retrospection	Why? Was it X that caused Y? What if I had acted differently?	Was it the aspirin that stopped my headache? Would Kennedy be alive had Oswald not shot him? What if I had not been smoking the past 2 years?

Figure 3: Three Levels of Causality (Pearl, 2019).

Most of our machine learning models are working on the first level, association. It means that "they invoke purely statistical relationships, defined by the naked data" (Pearl, 2019).

The second level is the intervention, which means that the algorithm relates to some possible interventions. It is not sufficient to see what it is, but also to change what we see. Actually, the problem is not simply $P(x|y)$. We need to consider $P(x|do(y))$. Here, $P(x|y)$ is observational conditional while $P(x|do(y))$ is intervention conditional. We take the barometric measurement as an example to see the difference between intervention and observation. Suppose y is the barometer reading and x is the actual atmosphere pressure, then $P(x|y)$ is almost 1 if the barometer is accurate. However, x is not determined on how we change the barometer reading. The atmospheric pressure will absolutely not be changed by the movement of the pointer of our barometer. Hence, $P(x|y) \neq P(x|do(y))$. In order to reach this level, Judea Pearl has discovered a method named "do-calculus", which can make the expressions not contains $do()$ and hence can be calculated by observing data (Tucci, 2013). Actually, SCM can also reach this level and we will discuss this later.

The highest level is the counterfactual. It can be expressed as the probability $P(y_x|x', y')$. It represents that if x' and y' happened, the probability of y of x' did not happen while x happened. Take a soccer game as an example. Suppose we replace player A with player B during the game and we win. Many may think that it is the replacement that changes the outcome of the game. The counterfactual is to figure out what would be the result if we did not replace A with B. Fortunately, the following introduced model, SCM, provides a method to solve this sort of problem.

2.3 Event Causality and Process Causality

In order to construct a causal model that can judge the causal relations, we need to distinguish the concepts of "Event Causality" and "Process Causality". Event causality is the relation of two events such as A and B, where A and B can be the coming of summer and forest fire here. As a contrast, if A and B are not events that happen at a point of time or processes, the causality relation between them is called process causality (Pearl, 2009b). For instance, consider the causality between the development of mathematics and computer science. We can say that it is the discovery of the mathematical theories that render the rapid development of computer science while computer science helped mathematicians to solve the problems like four-color theorem (Appel and Haken, 1976). It seems that the causality between them does not have a clear direction, which means that it is hard to define which one is the reason. However, there are still two methods to solve this sort of problem. One is to split the processes into many events. For the example above, we can divide the development of mathematics into some specific conclusions such as the discovery of Bayesian Formula which leads to Bayesian Network and the inventions of statistic methods which lead to most of the machine learning models nowadays. Also, we can also divide the development of computer science into some specifics algorithms. It is the specific algorithm that helps Kenneth Appel and Wolfgang Haken to solve the four-color theorem. Therefore, a process causality can be divided into some Event Causalities. Another is to use the model called Causal Loop Diagram in the fifth section.

The difference between event causality and process causality and their solutions are demonstrated in figure 4. In the next two sections, we will demonstrate a model designed for Event Causality called Structured Causal Model and its applications by now. In the fifth section, the Causal Loop Diagram designed for Process Causality will be introduced.

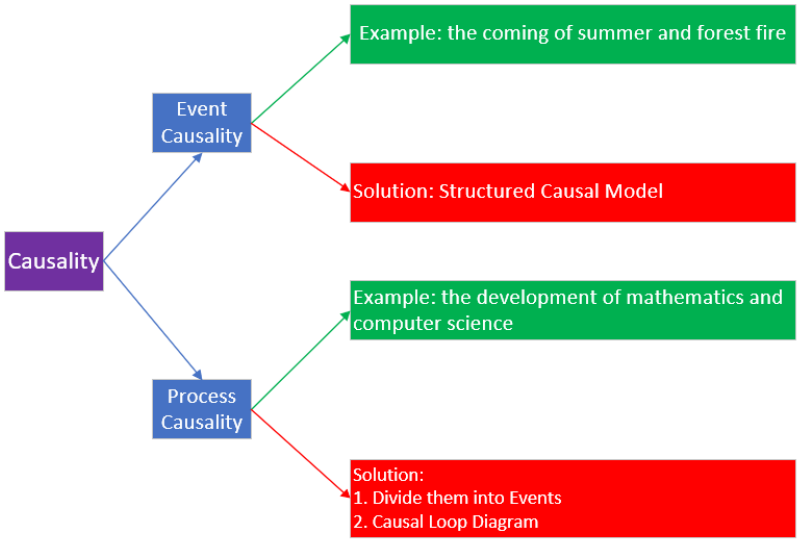

Figure 4: The Difference between Event Causality and Process Causality and Their Solutions.

3. From Bayesian Network to SCM

SCM is called Structured Causal Model. It is a mathematic framework widely applied in epidemiology, which is designed for event causality problems. Actually, it can be seen as the improvement of the Bayesian Network. To begin with, let us see how Bayesian Network works.

3.1 Bayesian Network

3.1.1 Bayes' Theorem

The inspiration of Bayesian Network has come from Bayes' Theorem. English statistician Thomas Bayes firstly discovered it in the essay titled "An Essay towards Solving a Problem in the Doctrine of Chances" (Bayes, 1763). It is a formula to calculate contingent probability. Suppose A and B are two events, $P(A|B)$ is the probability of A with B happens, then we have the contingent probability as below:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)} \tag{2}$$

What matters about Bayes' Theorem is that it has changed people's perception of probability. Different from the traditional opinion to study sample space, Bayes' Theorem provides an idea to compute the distribution of parameters. That is to say, the probability may not a fixed parameter if there is no sample and $P(A|B)$ is the probability after the observation of B, which is called posterior probability. Obviously, we don't have enough

samples when we are in face of new things and the samples are accumulated day by day so the perception of Bayes is more in line with nature (we need a distribution of probabilities at first and our knowledge will accumulate since we get the posterior probability).

3.1.2 The Structure of the Bayesian Network

The introduction is based on the essay titled Bayesian Network by Judea Pearl (Pearl, 2011) and the book named Causality (Pearl, 2009b). Bayesian Network is a DAG (Directed Acyclic Graph). Although it can only express correlation, it is the basis of SCM. Suppose $G = (I, E)$ is a DAG where $I = \{X_1, X_2, \cdots, X_n\}$ and E are the sets of nodes and directed edges, correspondingly. In a Bayesian Network, the nodes $\{X_1, X_2, \cdots, X_n\}$ represent the random variables and the directed edges represent two endpoints that are not conditional independent as shown in figure 5. What should be paid attention to is that compared with conventional ways to load the joint distribution model (the space complexity is $O(a^n)$), the space complexity of Bayesian Network is much lower ($O(an)$).

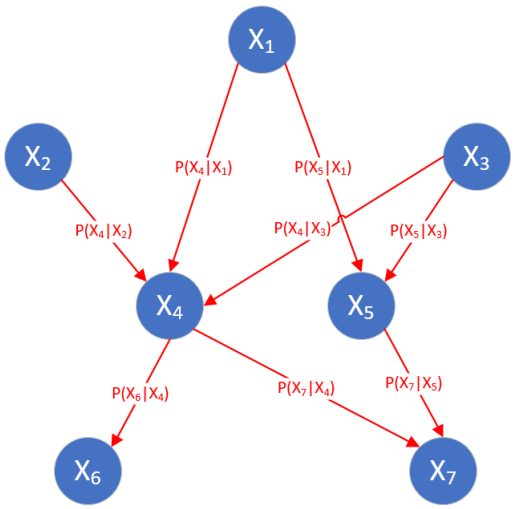

Figure 5: The Structure of Bayesian Network.

Suppose $pa(a)$ is the set of parent nodes of a. Then any node X stores $\{P(X|Y)|Y \in pa(X)\}$ which is the probability distribution of X under the condition of its parent node Y. Actually, the weight of the edge (Y, X) is $P(X|Y)$. Then the joint probability of any random variable can be computed as below:

$$P(X_1, X_2, \cdots, X_k) = P(X_k|X_1, X_2, \cdots, X_{k-1}) \tag{3}$$

$$\times P(X_{k-1}|X_1, \cdots, X_{k-2}) \times \cdots \times P(X_2|X_1) \times P(X_1) \tag{4}$$

There are three possible structures in a Bayesian Network. To simplify the notation, we use the vertical symbol to express X and Y are independent. The first one is called head-to-head (V-structure) as shown in figure 6. Then we have $(A \not\perp C)|B$ if B has been observed and the following results if B is unknown:

$$P(A, B, C) = P(A) \times P(C) \times P(B|A, C) \tag{5}$$
$$\Rightarrow P(A, C) = P(A) \times P(C) \tag{6}$$
$$\Rightarrow A \perp C \tag{7}$$

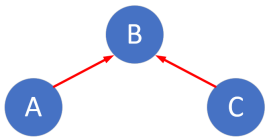

Figure 6: Head-to-Head.

The second one is tail-to-tail (common parent) as shown in figure 7.

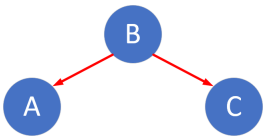

Figure 7: Tail-to-Tail.

We have the following results if B is observed and known. The structure is called tail-to-tail conditional independence.

$$P(A, C|B) = \frac{P(A, C, B)}{P(C)}, P(A, B, C) = P(B) \times P(A|B) \times P(C|B) \tag{8}$$
$$\Rightarrow P(A, C|B) = P(A|B) \times P(C|B) \tag{9}$$
$$\Rightarrow (A \perp C)|B \tag{10}$$

And if B is unknown, we have

$$P(A, B, C) = P(B) \times P(A|B) \times P(C|B) \tag{11}$$
$$\Rightarrow A \not\perp C \tag{12}$$

The third one is head-to-tail (cascade) as shown in figure 8.

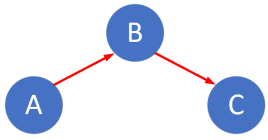

Figure 8: Head-to-Tail.

We have the following results if B is observed and known.

$$P(A, C|B) = \frac{P(A, B, C)}{P(B)}, P(A, B) = P(A) \times P(B|A) = P(B) \times P(A|B) \tag{13}$$
$$\Rightarrow P(A, C|B) = P(A|B) \times P(C|B) \tag{14}$$
$$\Rightarrow (A \perp C)|B \tag{15}$$

And if B is unknown, we have

$$P(A, B, C) = P(A) \times P(B|A) \times P(C|B) \tag{16}$$
$$\Rightarrow A \not\perp C \tag{17}$$

As a conclusion, a Bayesian Network consists of these three structures. This decomposition greatly simplifies the analysis of the network such as D-Separation (Geiger et al., 1990).

3.1.3 D-Separation

The full name of D-Separation is directed separation. The definition is that if a set of nodes O can separate A and B, if and only if there is no active path between A and B. If every three continuous nodes X, Y, Z in path P have one of the following properties, path P is an active path or A and B are not completely independent.

$$X \leftarrow Y \leftarrow Z, Y \notin O \tag{18}$$
$$X \rightarrow Y \rightarrow Z, Y \notin O \tag{19}$$
$$X \leftarrow Y \rightarrow Z, Y \notin O \tag{20}$$
$$X \rightarrow Y \leftarrow Z, Y \in O \tag{21}$$

If A and B are not D-Separated, we call they are a D-Connection. The conclusion can simplify the process of deciding whether two variables are independent or not.

3.1.4 An Experiment of Bayesian Network

Computation is not the key point here. Actually, there have already been methods like factor graph and Monte Carlo methods to compute the parameters in a Bayesian Network. However, an experiment for a simplified model will be demonstrated to show the importance of causality. Naive Bayesian is a classical algorithm for classifying. It can be seen as a Bayesian Network with no edges (Leung, 2007). It means that every two variables are independent therefore it has not considered the association between events. Although it can solve some problems indeed, the next experiment will show that the accuracy is not that satisfying and the limitations of the target problems. Actually, that is why we need to have Bayesian Network and further improvements such as SCM.

The program is a news classifier as shown in figure 9. Actually, that is nearly all the things that a Naive Bayesian Network can do. It can only classify objects that are

```
from sklearn.datasets import fetch_20newsgroups
from sklearn.model_selection import  train_test_split
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.naive_bayes import MultinomialNB
from sklearn.metrics import classification_report

news = fetch_20newsgroups(subset='all')
print(len(news.data))

X_train,X_test,y_train,y_test = train_test_split(news.data,news.target,test_size
print X_train[0]
print y_train[0:100]

vec = CountVectorizer()
X_train = vec.fit_transform(X_train)
X_test = vec.transform(X_test)

mnb = MultinomialNB()
mnb.fit(X_train,y_train)
y_predict = mnb.predict(X_test)

print('The Accuracy of Naive Bayes Classifier is:', mnb.score(X_test,y_test))
print(classification_report(y_test, y_predict, target_names = news.target_names)
```

Figure 9: News Classifier Based on Naive Bayesian.

independent enough such as gender classifying and text classifying. Actually, the accuracy is low as shown in figure 10 for the reason that there are still some associations in contents that have not been dug out. Human beings can understand the news but Naive Bayesian can only search for some specific independent features.

```
The Accuracy of Naive Bayes Classifier is: 0.8397707979626485
                          precision    recall  f1-score   support

             alt.atheism       0.86      0.86      0.86       201
           comp.graphics       0.59      0.86      0.70       250
 comp.os.ms-windows.misc       0.89      0.10      0.17       248
comp.sys.ibm.pc.hardware       0.60      0.88      0.72       240
   comp.sys.mac.hardware       0.93      0.78      0.85       242
          comp.windows.x       0.82      0.84      0.83       263
            misc.forsale       0.91      0.70      0.79       257
               rec.autos       0.89      0.89      0.89       238
         rec.motorcycles       0.98      0.92      0.95       276
      rec.sport.baseball       0.98      0.91      0.95       251
        rec.sport.hockey       0.93      0.99      0.96       233
               sci.crypt       0.86      0.98      0.91       238
         sci.electronics       0.85      0.88      0.86       249
                 sci.med       0.92      0.94      0.93       245
               sci.space       0.89      0.96      0.92       221
  soc.religion.christian       0.78      0.96      0.86       232
      talk.politics.guns       0.88      0.96      0.92       251
   talk.politics.mideast       0.90      0.98      0.94       231
      talk.politics.misc       0.79      0.89      0.84       188
      talk.religion.misc       0.93      0.44      0.60       158

                accuracy                           0.84      4712
               macro avg       0.86      0.84      0.82      4712
            weighted avg       0.86      0.84      0.82      4712
```

Figure 10: The Experiment Result of Naive Bayesian.

3.2 Structural Causal Model

In a Bayesian Network, $A \to B \to C$ and $A \leftarrow B \leftarrow C$ are the same without a doubt. However, causality is directed so Bayesian Network still loses some information about causality. Actually, it can not express intervention so some improvements are needed. This subsection will demonstrate the structure of SCM and the next section will introduce how it works.

3.2.1 Rubin Causal Model

RCM (Rubin Causal Model) was put forward by a statistician Donald Rubin from Tsinghua University (Sekhon, 2008). Suppose U is the set of event systems u and u can be any events or objects that we are observing. T is the set of interventions t that have impacts on u. Y is the state function. For example, for system u and intervention t, the state y is $y = Y_t(u)$. It is worth noticing that there exists a $c \in T$ such that c is the empty intervention, which means that doing nothing to the system. In fact, c can be defined in accordance with requirements. Then in RCM model, the causality is defined by $\delta(u) = Y_t(u) - Y_c(u)$. There is a practical problem in this model. It is called the Fundamental Problem of Causal Inference (Holland, 1986), which means that we cannot get $Y_t(u)$ and $Y_c(u)$ simultaneously. The solution is to establish some extra hypotheses. They are SUTVA (Stable Unit Treatment Value Assumption), Assumption of Constant Effect and Assumption of homogeneity.

However, there is still a problem. If we add some parameters or enlarge the volume of the model, the training time and training data will increase exponentially. Luckily, this is a problem that can be solved by the structure of the Bayesian Network.

3.2.2 The Structure of SCM

We are given a set of observables $X_1, \cdots, X_n$ (modeled as random variables) associated with the vertices of a DAG (Directed Acyclic Graph) G. We assume that each observable is the result of an assignment

$$X_i = f_i(pa(X_i), U_i) \tag{22}$$

using a deterministic function f_i depending on X_i's parents in the graph (denoted by $pa(X_i)$) and on a stochastic unexplained variable U_i (Schölkopf, 2019). The directed edges in the graph represent causality. Here, $pa(X_i)$ is the endogenous variable while U_i is the exogenous variable. The structure of SCM is as shown in figure 11. "The noise U_i ensures that the overall object (22) can represent a general conditional distribution $p(X_i|pa(X_i))$, and the set of noises $U_1, \cdots, U_n$ are assumed to be jointly independent" (Schölkopf, 2019). Contrasting SCM with RCM, we can see that the space complexity of SCM is much lower (benefit from Bayesian Network) and the structure is much clearer to see the pilot process while we can only get the reason and cause without pilot process in RCM.

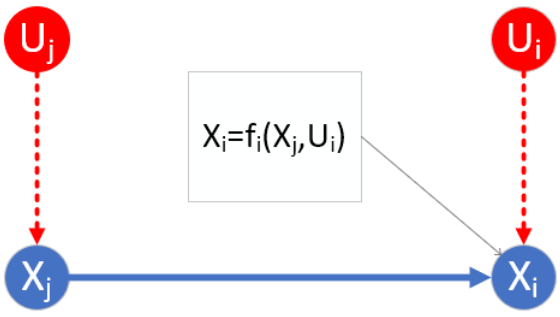

Figure 11: The Structure of SCM.

4. Properties and Applications of SCM

In this section, we will introduce how can SCM reach the causality level of intervention and counterfactual. With these properties, SCM provides various tools to solve various problems, which will also be mentioned in the following pages.

4.1 Properties

4.1.1 Intervention

As we mentioned above, Bayesian Network can solve problems like $P(Y|E=e)$, where Y is the unknown variables and e is what we observed. SCM can actually solve the intervention problems like $P(Y|E=e, do(X=x))$. Take X_i, X_j, X_k as an example, suppose they are continuous nodes in an SCM as shown in figure 12.

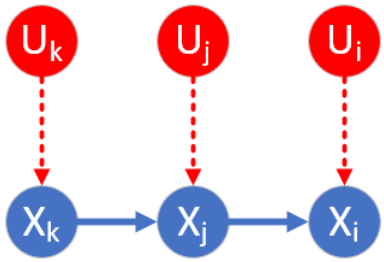

Figure 12: X_i, X_j, X_k in a SCM.

Then the system can be expressed by the equations

$$X_k = f_k(U_k) \tag{23}$$
$$X_j = f_j(X_k, U_j) \tag{24}$$
$$X_i = f_i(X_j, U_i) \tag{25}$$

before the intervention done on X_j. When the intervention $do(X_j = x)$ is done, the model will eliminate the edge between X_k and X_j as it shown in figure 13 and the system can be expressed as below

$$X_k = f_k(U_k) \tag{26}$$
$$X_j = x \tag{27}$$
$$X_i = f_i(X_j, U_i) \tag{28}$$

According to Markov's Theorem, we have $P(X_1, \cdots, X_n) = \prod_{i=1}^{n} P(X_i|pa(X_i))$. Then after an intervention $do(X=x)$. The equation is as below:

$$P(X_1, \cdots, X_n|do(X=x)) = \prod_{i=1, X_i \notin X}^{n} P(X_i|pa(X_i))|_{X=x} \tag{29}$$

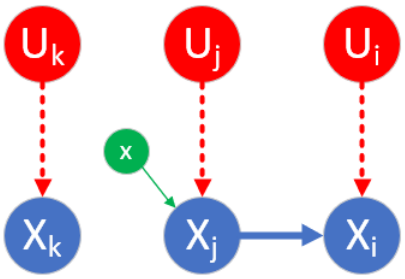

Figure 13: X_i, X_j, X_k after the Intervention $do(X_j = x)$.

In fact, there is a more generalized equation (Pearl, 2009b)

$$P(Y = y|do(X = x)) = \sum_t P(Y = y|T = t, X = x)P(T = t) \tag{30}$$

which can simplify the computing process of SCM. It shows that interventions will not break the Markov's condition. Simultaneously, back-door criterion can also optimize the computation in SCM according to MH Maathuis and D Colombo (Maathuis et al., 2015).

4.1.2 Counterfactual

Recall that the counterfactual conditional probability can be expressed as $P(y_x|x', y')$. Actually, x' and y' are not symmetric because x' is the intervention and y' is the result. Suppose $E()$ express the causality. Then the causality between intervention and result is $E(Y_x|X = x')$, which can be computed by SCM. Actually, Judea Pearl came up with another idea which is to compute $E(Y'_x|X = x)$. It is called ETT (Effect of Treatment on the Treated). ETT can be used in machine learning, which is called intent-specific optimization by Judea Pearl (Forney et al., 2017). More about counterfactual has been introduced in A Framework for Empirical Counterfactuals, or for All Intents, a Purpose by A Forney (Forney, 2018).

4.2 Applications

4.2.1 Cause-Effect Discovery

One way to realize cause-effect discovery is by using Markov's condition (equation (29)). However, in actual life, the data set is limited so conditional independence tests are very complicated. Therefore, this method is inefficient and useless. In fact, we can add some limitations to the function in SCM in order to simplify the training process. Suppose there is an SCM with two nodes X, Y. Then we have

$$X = U \tag{31}$$
$$Y = f(X, V) \tag{32}$$

and $U \perp V$. Suppose $V \in F = \{f_v(x) \equiv f(x, v)|v \in supp(V)\}$, then because $f(x, v)$ depends on v, it is extremely hard to get information from the data set (V is not observable).

Therefore, the additive noise model is considered (Schölkopf, 2019).

$$X = U \tag{33}$$
$$Y = f(X) + V \tag{34}$$

If there is no such limitation, the function may be not linear and the computation complexity will be exponential. In Shohei Shimizu's work, there is a method called LiNGAM for causal discovery and it is under the assumption that the data generating process is linear (Shimizu et al., 2006). It relies on independent component analysis so it does not require pre-specified time-ordering of the variables. Simultaneously, there is some work done on the nonlinear processes. PO Hoyer pointed out that many causal relationships are more or less nonlinear, raising some doubts as to the applicability and usefulness of purely linear methods (Hoyer et al., 2009). His model is based on using the nonlinearities as a blessing instead of a curse. "They typically provide information on the underlying causal system and allow more aspects of the true data-generating mechanisms to be identified " (Hoyer et al., 2009). Recent progress is done by K Chalupka. He came up with FIT (Fast Independence Test). The basic idea is that when $P(X|Y, Z) = P(X|Y)$, Z is not useful. However, while $P(X|Y, Z) \neq P(X|Y)$, Z might improve prediction results (Chalupka et al., 2018). Actually, the recent works are mostly done by kernel function classes.

4.2.2 Robust Learning

As introduced in Robust Learning via Cause-Effect Models, knowledge of an underlying causal direction can facilitate some tasks including covariate shift, concept drift, transfer learning and semi-supervised learning (Schölkopf et al., 2011). What renders the possible optimization is the properties of cause and effect we mentioned above. In the first part of the essay, the writer introduces some properties of SCM. Most of them are mentioned above. In the second part and third part, the author introduces some methods of robust learning of predicting cause from effect and the opposite direction. For example, for the latter case, there is some additional information about the output for semi-supervised learning. We are given training points sampled from $P(X, Y)$ and an additional set of outputs sampled from $P(Y)$. The goal is to estimate $P(Y|X)$. Then the additive noise model can be used and the assumption can make it to a weaker problem (Schölkopf et al., 2011). This paper shows that the cause-effect model can be very useful for some special problems and simultaneously, robustness to covariate shift.

Actually, another paper solved a problem as below (Daniusis et al., 2012). If an SCM is $X \to Y$ and we need to learn it. The decomposition is $P(X, Y) = P(X)P(Y|X)$ so according to ICM, $P(X)$ should not have any information in $P(Y|X)$. Hence it is unlikely possible to use SSL. However, the opposite direction is possible. It is a possible method for SCM to improve machine learning.

Except for SSL, SCM can also solve the problem of adversarial vulnerability. The main idea is to decompose the module (usually, the neural network) into some components. These components are corresponding to the causalities. Then the module will have some robustness because of the robustness of causality. More specifically, we can use the following

factorization to do the decomposition.

$$p(X_1, \cdots, X_n) = \prod_{i=1}^{n} p(X_i | X_{i+1}, \cdots, X_n) \tag{35}$$

"It was shown that a possible defense against adversarial attacks is to solve the anti causal classification problem by modeling the causal generative direction" (Schölkopf, 2019). Actually, in the work done by Lukas Schott, he has already used analysis by synthesis to construct a robust classification model (Schott et al., 2018). The novel classification model works well on MNIST, especially on solving the problem of adversarial vulnerability.

Meanwhile, SCM can solve the problems of disturbance of reinforcement learning which is much closer to cause-effect studying (Lu et al., 2018). The paper considers the problem of learning from both observed actions and rewards. As the author says, "this is the first time that confounders are taken into consideration for addressing full RL problems with observational data" (Lu et al., 2018), the idea of SCM may make a bigger donation to RL in the future.

4.2.3 Other Applications

According to Causality for Machine Learning, SCM can also play a role in learning transferable mechanisms (Schölkopf, 2019). Take the example of fish again, it is not because we see a fish in the grass many times that we can recognize it. It is because of the mechanisms in our minds. Traditional machine learning will not have enough data in the real world but SCM can help with the learning of transferable mechanisms. Another application is to express the interventional world model. This will be an important approach to strong AI. But the details of this subject are somehow closer to philosophy, which will not be discussed here.

4.2.4 Applications in Real World

In this part, we will demonstrate an application of SCM and see how can it work in the real world. The paper A Comparison of Approaches to Advertising Measurement: Evidence from Big Field Experiments at Facebook is actually a measurement of the causal effects of digital advertising (Gordon et al., 2019). The author mainly uses analysis of RCT to weight the effectiveness of the advertising places and compares them with some typical observational methods. The most important part is the cause effects in the RCT. Actually, cause effects play a crucial role in advertising, which may not be detected by observational methods instead of experiments. And we can see the results from RCT as shown in figure 14. However, the results from observational methods usually overestimate the results of RCT as shown in figure 15. The bias may be very large. The fifteen studies in the paper show the conclusion that the methods they study yield biased estimates of causal effects of advertising in a majority of cases. As the paper says, it is also measurement of observational versus experimental approaches to causality. It indicates that the research of cause effects are not merely theoretic. Actually, it indeed has some practical uses.

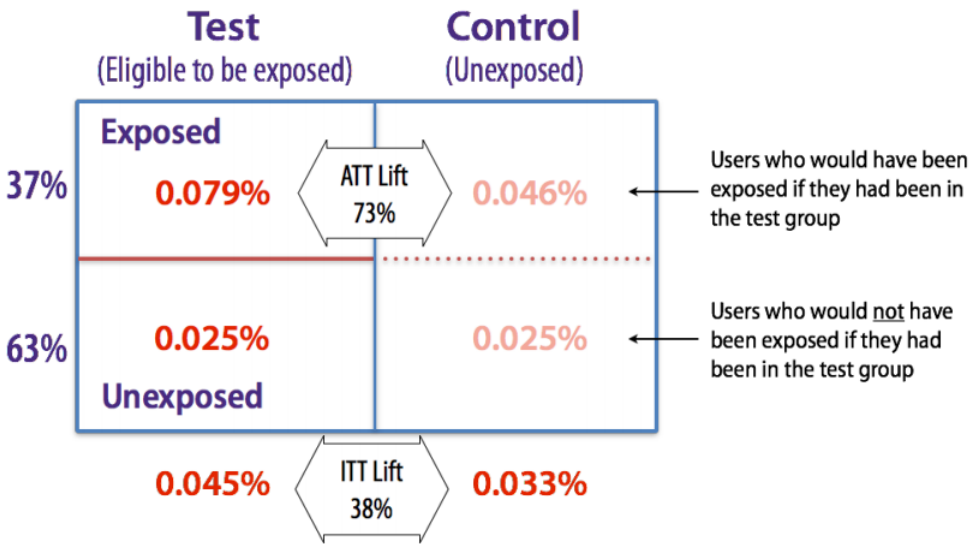

Figure 14: Result from RCT (Gordon et al., 2019).

5. A Simple Approach To CLD

Different from the methods based on event causality, CLD (Causal Loop Diagram) is a model to solve the problems of procedure causality. For example, for A and B, they can both last for a period of time and they can be reciprocal causation. The relation can not be expressed by SCM for the reason that the edges in SCM are directed and SCM does not allow edges like $A \leftrightarrow B$ (or $A \rightarrow B$ and $A \leftarrow B$ at the same time). The diagram shown in figure 15 is a typical instance of CLD.

Suppose the relation between A and B can be expressed as $A = \alpha B + \beta$. Then if $\alpha = \frac{\partial A}{\partial B} > 0$, we claim that B has a positive influence on A. If $\alpha = \frac{\partial A}{\partial B} < 0$, we claim that B has a negative influence on A. There are some definitions to research this sort of problems. If $A \rightarrow B$ and $B \rightarrow A$ are all positive or negative, we call it a reinforcement feedback loop. If else, we call it a balanced feedback loop. In fact, in a loop, suppose there are n negative influences. If $n \equiv 1(mod 2)$, the loop is a balanced feedback loop such as the CLD in figure 15. If $n \equiv 0(mod 2)$, the loop is a reinforcement feedback loop, for example, the compound interest of banks. More properties of CLD can be found in Problems with causal-loop diagrams by George P. Richardson (Richardson, 1986).

6. Future Development of Causality in AI

In conclusion, causality in machine learning is under developing now and there are many problems waiting to be solved in the future. The first problem is how to find effective methods to compute SCM. Nowadays, we can only solve the problems with strict constrains and there are actually many more problems in real life. The model of causality comes from economics and it is a basic content in the researches of economics. We can not just indulge in curve fitting. We should find ways to make better use of our computation resources.

		(A)	(C)	(D)	(E)	(F)	(G)	(H)	(I)	(J)	(K)	(L)	(M)	(N)	(O)
			EM	Propensity Score Matching				Regression Adjustment				Stratified Regression			
Campaign	Outcome	RCT Lift*	Age, Gender	Age, Gender + FB Vars	Age, Gender + FB Vars + Census Vars	Age, Gender + FB Vars + Census Vars + Activity Vars	Age, Gender + FB Vars + Census Vars + Activity Vars +FB Match Vars	Age, Gender + FB Vars	Age, Gender + FB Vars + Census Vars	Age, Gender + FB Vars + Census Vars + Activity Vars	Age, Gender + FB Vars + Census Vars + Activity Vars +FB Match Vars	Age, Gender + FB Vars	Age, Gender + FB Vars + Census Vars	Age, Gender + FB Vars + Census Vars + Activity Vars	Age, Gender + FB Vars + Census Vars + Activity Vars +FB Match Vars
1	Checkout	30%	116%	109%	107%	85%	93%	104%	99%	88%	76%	101%	94%	65%	51%
2	Checkout	1.3%	432%	161%	149%	37%	36%	149%	140%	43%	35%	97%	98%	54%	40%
3	Checkout	8.8%	65%	20%	24%	41%	17%	21%	23%	38%	5%	18%	19%	30%	2%
4	Checkout	73%	222%	145%	131%	143%	95%	126%	122%	134%	100%	98%	87%	96%	74%
5	Checkout	450%	511%	418%	443%	463%	316%	428%	432%	437%	305%	447%	431%	435%	301%
7	Checkout	2.7%	37%	20%	18%	-33%	-36%	19%	20%	-33%	-35%	19%	19%	-31%	-33%
8	Checkout	-2.9%	48%	31%	36%	50%	27%	36%	41%	54%	29%	32%	37%	52%	28%
9	Checkout	2.4%	3414%	2062%	1970%	2314%	1710%	1994%	1999%	2319%	1716%	1962%	1962%	2210%	1656%
10	Checkout	2.0%	38%	23%	16%	43%	-7%	20%	20%	34%	-13%	21%	21%	35%	-11%
11	Checkout	9%	275%	29%	31%	38%	7%	30%	31%	35%	3%	30%	31%	34%	2%
12	Checkout	1%	129%	111%	110%	82%	82%	112%	111%	82%	81%	112%	111%	84%	82%
13	Checkout	-15%	-39%	-35%	-36%	-30%	-31%	-35%	-35%	-31%	-30%	-35%	-35%	-31%	-30%
14	Checkout	62%	119%	80%	85%	95%	101%	80%	83%	92%	90%	74%	77%	82%	84%
15	Checkout	2%	26%	-10%	-9%	-10%	-13%	-9%	-9%	-11%	-14%	-9%	-9%	-12%	-14%
1	Registration	781%	1024%	978%	944%	1060%	977%	968%	960%	1087%	985%	824%	800%	432%	348%
5	Registration	893%	1270%	1071%	1055%	1070%	765%	1067%	1067%	1063%	728%	1112%	1104%	1081%	772%
8	Registration	63%	180%	162%	159%	173%	167%	150%	153%	158%	114%	157%	161%	160%	125%
10	Registration	9%	34%	19%	18%	34%	-3%	18%	18%	31%	0%	19%	18%	31%	2%
14	Registration	158.1%	275%	215%	219%	244%	241%	219%	219%	238%	234%	219%	218%	240%	239%
2	Page View	1517%	4261%	2493%	2416%	1150%	1177%	2408%	2422%	1175%	1187%	1162%	1181%	1722%	1268%
5	Page View	609%	846%	771%	731%	719%	484%	751%	748%	710%	477%	776%	769%	717%	498%
6	Page View	14%	227%	103%	105%	263%	255%	103%	106%	250%	246%	111%	115%	255%	278%

* Red: RCT Lift is statistically different from 0 at 5% significance level

Observational method overestimates lift

Observational method underestimates lift

Color proportional to overestimation factor; darkest color reached at 3-times over- or underestimation

Figure 15: Summary of lift results (Gordon et al., 2019).

SCM may not the final answer but it still gives us a train of thought to add causality to machine learning. In order to reach the level of strong AI, causality has to be attached importance to.

Another problem is that we still lack comprehension of the causality. More specifically, we still can not understand the way by which animals learn. The models including interventions and counterfactuals can not get free will, which is the core problem of artificial intelligence (Pearl, 2009a).

More practically, today's machine learning still contains insufficient elements of causality. In the near future, we may make some progress on the applications of cause-effect models. If the applications are convincing, the topic will naturally turn to higher levels of causality. For example, it may help to solve disentangled representation in the near future. Some works have already been done (Besserve et al., 2018) but not convincing enough to change most of the people's attitudes towards the future of causality in machine learning.

Judea Pearl has predicted the development of causality in AI, "the first step, one that will take place in maybe 10 years, is that conceptual models of reality will be programmed by humans. The next step will be that machines will postulate such models on their own and will verify and refine them based on empirical evidence" (Hartnett, 2018).

Just as he said, the future of AI can not be a curve fit. It is causality that can bring the ability of machine learning to a brand new level.

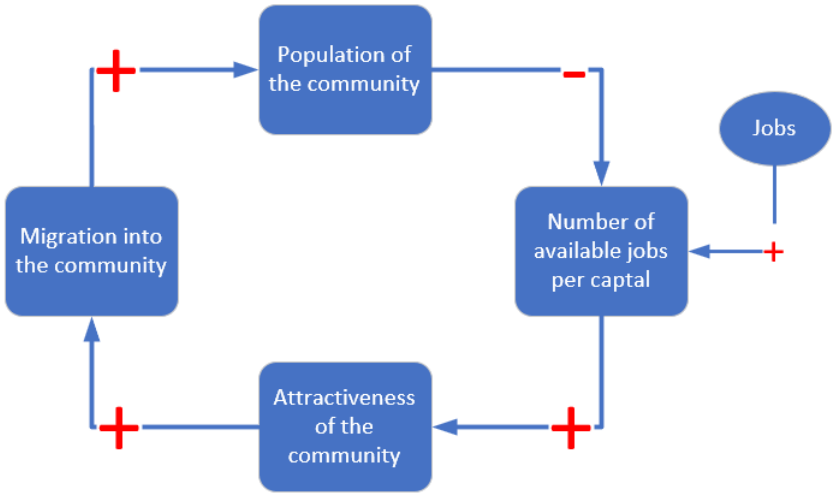

Figure 16: A Typical Example of CLD.

References

Kenneth Appel and Wolfgang Haken. Every planar map is four colorable. *Bulletin of the American mathematical Society*, 82(5):711–712, 1976.

Thomas Bayes. Lii. an essay towards solving a problem in the doctrine of chances. by the late rev. mr. bayes, frs communicated by mr. price, in a letter to john canton, amfr s. *Philosophical transactions of the Royal Society of London*, (53):370–418, 1763.

Michel Besserve, Rémy Sun, and Bernhard Schölkopf. Counterfactuals uncover the modular structure of deep generative models. *arXiv preprint arXiv:1812.03253*, 2018.

Krzysztof Chalupka, Pietro Perona, and Frederick Eberhardt. Fast conditional independence test for vector variables with large sample sizes. *arXiv preprint arXiv:1804.02747*, 2018.

Povilas Daniusis, Dominik Janzing, Joris Mooij, Jakob Zscheischler, Bastian Steudel, Kun Zhang, and Bernhard Schölkopf. Inferring deterministic causal relations. *arXiv preprint arXiv:1203.3475*, 2012.

AP Dempster. Causality and statistics. *Journal of statistical planning and inference*, 25(3): 261–278, 1990.

Georges Dicker. *Hume's epistemology and metaphysics: an introduction*. Routledge, 2002.

Andrew Forney. *A Framework for Empirical Counterfactuals, or for All Intents, a Purpose*. PhD thesis, University of California, Los Angeles, 2018.

Andrew Forney, Judea Pearl, and Elias Bareinboim. Counterfactual data-fusion for online reinforcement learners. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the*

34th International Conference on Machine Learning, volume 70 of *Proceedings of Machine Learning Research*, pages 1156–1164, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR. URL http://proceedings.mlr.press/v70/forney17a.html.

Dan Geiger, Thomas Verma, and Judea Pearl. d-separation: From theorems to algorithms. In *Machine Intelligence and Pattern Recognition*, volume 10, pages 139–148. Elsevier, 1990.

Brett R Gordon, Florian Zettelmeyer, Neha Bhargava, and Dan Chapsky. A comparison of approaches to advertising measurement: Evidence from big field experiments at facebook. *Marketing Science*, 38(2):193–225, 2019.

Kevin Hartnett. To build truly intelligent machines, teach them cause and effect. *Internet: https://www.quantamagazine.org/to-build-truly-intelligent-machines-teach-them-cause-and-effect-20180515/*, 2018.

Paul W Holland. Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960, 1986.

Patrik O. Hoyer, Dominik Janzing, Joris M Mooij, Jonas 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. Curran Associates, Inc., 2009. URL http://papers.nips.cc/paper/3548-nonlinear-causal-discovery-with-additive-noise-models.pdf.

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. *arXiv preprint arXiv:1905.02175*, 2019.

K Ming Leung. Naive bayesian classifier. *Polytechnic University Department of Computer Science/Finance and Risk Engineering*, 2007.

Chaochao Lu, Bernhard Schölkopf, and José Miguel Hernández-Lobato. Deconfounding reinforcement learning in observational settings. *arXiv preprint arXiv:1812.10576*, 2018.

Marloes H Maathuis, Diego Colombo, et al. A generalized back-door criterion. *The Annals of Statistics*, 43(3):1060–1088, 2015.

Dominique Drouet Mari and Samuel Kotz. *Correlation and dependence*. World Scientific, 2001.

Judea Pearl. Giving computers free will. *Internet: https://www.forbes.com/2009/06/18/computers-free-will-opinions-contributors-artificial-intelligence-09-judea-pearl.html*, 2009a.

Judea Pearl. *Causality*. Cambridge university press, 2009b.

Judea Pearl. Bayesian networks. 2011.

Judea Pearl. The seven tools of causal inference, with reflections on machine learning. *Commun. ACM*, 62(3):54–60, 2019.

George P Richardson. Problems with causal-loop diagrams. *System dynamics review*, 2(2): 158–170, 1986.

Bernhard Schölkopf. Causality for Machine Learning. *arXiv e-prints*, art. arXiv:1911.10500, Nov 2019.

Bernhard Schölkopf, Dominik Janzing, Jonas Peters, and Kun Zhang. Robust Learning via Cause-Effect Models. *arXiv e-prints*, art. arXiv:1112.2738, Dec 2011.

Lukas Schott, Jonas Rauber, Matthias Bethge, and Wieland Brendel. Towards the first adversarially robust neural network model on mnist. *arXiv preprint arXiv:1805.09190*, 2018.

Jasjeet S Sekhon. The neyman-rubin model of causal inference and estimation via matching methods. *The Oxford handbook of political methodology*, 2:1–citation_lastpage, 2008.

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

Robert R Tucci. Introduction to judea pearl's do-calculus. *arXiv preprint arXiv:1305.5506*, 2013.