Need For Causality Recognition When Data Mining In An Inherently Ill Defined World

Lawrence J. Mazlack

Applied Artificial Intelligence Laboratory
University of Cincinnati
Cincinnati, Ohio

mazlack@uc.edu

- World is inherently ill defined

- In *common-sense* lives, need to deal with:
    - imprecision
    - uncertainty
    - incompleteness
    - specificity lack
    - imperfect knowledge

- World is inherently ill defined

- In *common-sense* lives, need to deal with:
  - imprecision
  - uncertainty
  - incompleteness
  - specificity lack
  - imperfect knowledge

- Causality has central position in human reasoning

- Have *common-sense* belief in *causal*
  - existence
  - occurrence

- Essential in *common-sense* human decision making

- Difficult to *precisely* describe; yet we use it

> *"It is better to be approximately right than to be precisely wrong"* ... J. M. Keynes

- Recognizing cause/effect tantalizing goal throughout history

- People are *realists*, believe that the world is as perceived; i.e., are perception based

- Important to data mining:

  Understanding whether deterministic/coincidental relationship exists

  ⇒ decision value

  ⇑

  primary reason
  for
  data mining

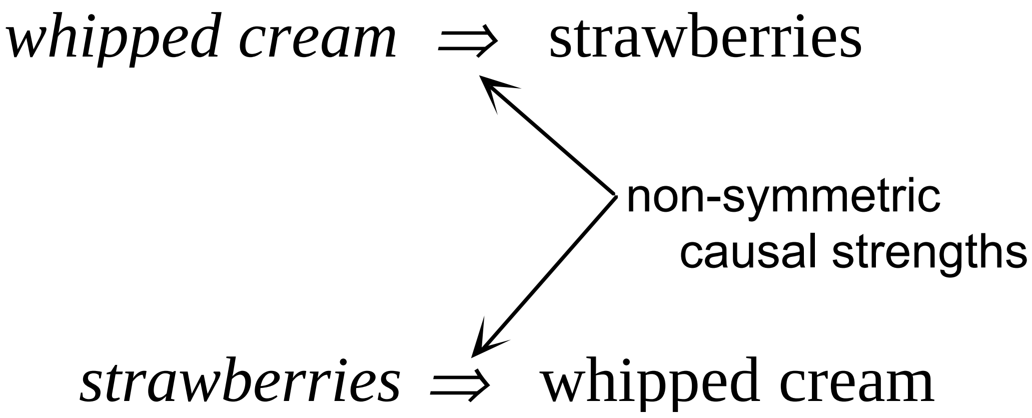
- *Sometimes*, data items may often occur together but may not have deterministic relationship

  Example:
  Shopper may buy both:
  - *bread*
  - *milk*

  Mostly,
  - *Milk* buy not caused by *bread* buy
  - *Bread* buy not caused by *milk* buy

  ⇒ No decision value in exclusively joint occurrence knowledge

- Alternatively, there may be causal effects
  - *Strawberries* buy *may* affect *whipped cream* buy purchase
  - *Some* people who buy *strawberries*
    - > want *whipped cream* with them
    - > *desire intensity* for *whipped cream* varies
  - Conditional primary effect: *whipped cream* modified by secondary effect: *desire*
  - Not necessarily symmetric causal strengths



whipped cream ⇒ strawberries

non-symmetric causal strengths

strawberries ⇒ whipped cream

- Data mining extracts information from data

- Most common methods build rules; *association* rules most common

Conditional rule:

IF Age < 20
THEN Income < $10,000
with {belief = 0.8}

Association rule:

Customers who buy beer and sausage also tend to buy mustard
with {confidence = 0.8}
{support = 0.15}

- In many ways, interest in extracted rules:

  Promise (or illusion) of causal, or at least, predictive relationships.

Association rule:

Customers who buy beer and sausage
also tend to buy mustard
with {confidence = 0.8}
{support = 0.15}

- At first glance:
  - Seem to imply cause/effect
    i.e., buying *beer* and *sausage*
    causes *mustard* purchase

- Only *existence* of statistical relation discovered
  - Not nature of relationship
  - Not known:
    > What causes what
    > Whether unspecified factor involved
    > Whether causality at all
  - Only joint occurrence frequency

- *Causal* relationships have greatest *decision value*

Naïve use association rule use can lead to trouble

- *Sometimes:*
  - causal relationships might be meaningful while
  - associations are not useful

Example: At store, customers buy:

- *hamburger 33%* of the time
- *hot dogs* 33% of the time
- both *hamburger* and *hot dogs* 33% of the time
- *sauerkraut* only if *hot dogs* are also purchased

	hamburger	hot dog	sauerkraut
t_1	1	1	1
t_2	1	0	0
t_3	0	1	1

Associations:

- (hamburger, hot dog) = 0.5
- (hamburger, sauerkraut) = 0.5
- (hot dog, sauerkraut) = 1.0

General:

- *sauerkraut = 66%*
- *hamburger = 66%*
- *hot dog = 66%*

Associations:

- (hamburger, hot dog) = 0.5
- (hamburger, sauerkraut) = 0.5
- (hot dog, sauerkraut) = 1.0

General:

- *sauerkraut = 66%*
- *hamburger = 66%*
- *hot dog = 66%*

If the merchant:

- Reduced price of hamburger (as sale item)
- Raised price of sauerkraut to compensate (as rule *hamburger* ⇔ *sauerkraut* has high confidence.
- Pricing compensation would not work:
  - *Sauerkraut* sales would not increase with *hamburger* sales
  - Most likely, sales of *hot dogs* (& consequently, *sauerkraut)* would likely decrease as buyers would substitute *hamburger* for *hot dogs*

- *Causal* relationships exist in *common-sense* world

  For example:
  If someone fails to stop at red light and
  - automobile accident happens, *can be said:*
    - failure to stop was the accident’s cause

- Another *common-sense* way to think of causal relationships is *counterfactually*:

For example:

If driver dies in accident:
- had accident *not* happened
  ⇒ would *not* be dead

- *False recognition* complicates causal recognition:
  - Coach may win game when wearing particular pair of socks, then always wear same socks
  - More interesting: *false causality* between: *music* and *motion*

    example, Lillian Schwartz:

    > Acquired series of computer generated random images

    > Sequenced them, and attached a sound track (usually Mozart).

    > Music was not matched to images;

    > However, when viewing, images and music appeared to be well connected

    > All connections were observer supplied

- Appear to be inherent limits on whether causality can be determined:

  - *Quantum Physics:*

    Heisenberg's uncertainty principle:

    Cannot precisely measure both particle's momentum and precision

    If something not observed, might not happen. If it was observed, probably changed both event and observer

  - *Incompleteness:*

    World <u>knowledge</u> might never be *complete* because we, as observers, are integral parts of what observed

  - *Gödel's Theorem:*

    Showed in any logical formulation of arithmetic that there would always be statements whose validity was indeterminate.

    Suggests that there will always be inherently unpredictable aspects of the future

  - *Turing Halting Problem:*

    Any problem solvable by a step-by-step procedure can be solved using Turing machine. However, many routines where cannot ascertain if Turing machine will take finite, or infinite number of steps.

    Curtain between what can and cannot be known mathematically

- Appear to be inherent limits on whether causality can be determined:
  - *Quantum Physics*
  - *Incompleteness*
  - *Gödel’s Theorem*
  - *Turing Halting Problem*

  - Chaos Theory:

    Chaotic systems appear to be deterministic;
    but are computationally irreducible

    If nature is chaotic, it might be fully deterministic,
    yet wholly unpredictable

  - *Space Time:*

    Einstein’s space time: What is “now” and “later” is local to observer; another observer may have contradictory views

    Cannot distinguish between: cause/effect by time line

  - *Arithmetic Indeterminism:*

    Arithmetic has random aspects that introduce uncertainty as to whether equations may be solvable.

    Chatin discovered that Diophantine equations may or may not have solutions, depending on parameters chosen. Whether parameter leads to solvable equation appears to be random. (Diophantine equations represent well-defined problems, emblematic of simple arithmetic procedures.)

Completeness major problem

- Events often affected by large number of potential factors

  example:
  With plant growth, many factors can all affect plant growth:

  temperature

  chemicals in the soil

  types of creatures present, etc.,.

- Unknown:
  - Identity of all possible factors
  - Relative importance of all factors
  - What causal factors will or will not be present
  - How many can be discouvered

⇒*Imperfect* knowledge of causal structure

- Given determinism’s uncertainty & imprecision,
  - *Precise* and *complete* knowledge of causal events is *uncertain*

- **BUT**, we have a common-sense belief that causal effects exist in real world

- **IF** can develop models tolerant of imprecision
  - Would be useful
  - Perhaps, soft computing tools may be useful

- Common sense cause-effect is not binary:
  - *When α occurs, β always occurs*

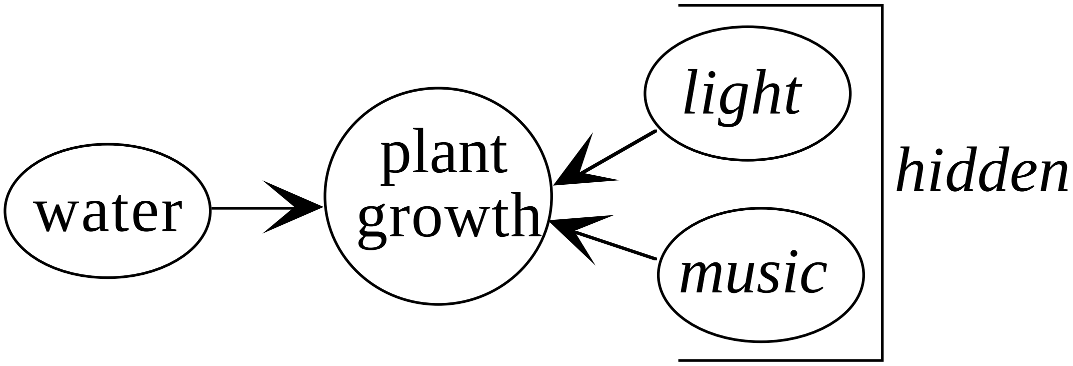
- Inconsistent with our common-sense understanding
  - Less complex:

    When champagne bottle hits ship,
    bottle usually breaks

  - More complex:

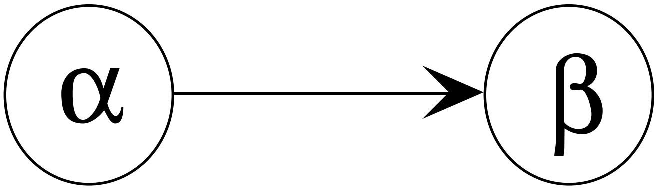
    When plant receives water,
    plant usually grows

- Some problems with discouvering causality:
  - - Adequate definition of a causal relation,
  - - Representation
  - - Computing causal strengths
  - - Missing (hidden) attributes having causal effect



  - - Distinguishing between:
    association & *causal* values,
  - - Inferring causes & effects from representation

- Dependent or causal relationship



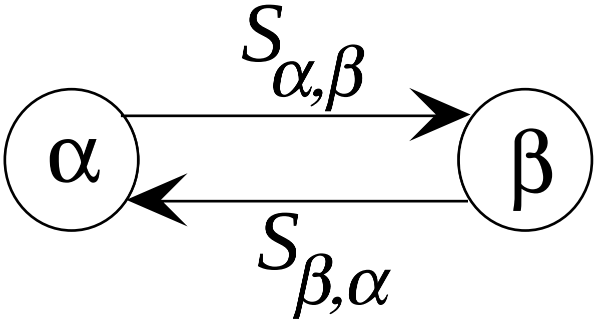
- To what degree does α cause β?

- Value for β proportionally sensitive to:
small change in the value of α?

- Always hold in time and in every situation?
If does not *always* hold,
can situation when it does hold be discouvered?

- How describe causal relationships?
> *degree* of causal strength (non-Boolean)?
> probability?
> possibility?

- Mutual dependencies; i.e., $\alpha \rightarrow \beta$ as well as $\beta \rightarrow \alpha$?



- Possible different strengths?

  α - *short men*

  β - *tall women*.

  $S_{\alpha,\beta}$ - social contact desire *usually* caused in: *short men* by the sight of *tall women*

- *usually:* $S_{\alpha,\beta} > S_{\beta,\alpha}$

In market basket data

- Associations find symmetric co-occurrence ratios

- Causals may be imbalanced dependencies

  - If customer first buys *strawberries*, good chance also buys *whipped cream*

  - Conversely, if first buys *whipped cream*,

    subsequent purchase of *strawberries* may be less likely;

    maybe more likely buy *ice cream* or *pie*

- Part of the difficulty of recognizing causality comes from identifying relevant data
  - Some data might be redundant, some irrelevant
  - Some more important than others.
  - Data can have a high dimensionality with only relatively few utilitarian dimensions
    - Data may have higher dimensionality than necessary to fully describe situation.
    - Data complexity may be unknown
    - Dimensionality reduction an important issue

Types Of Apparent Causality

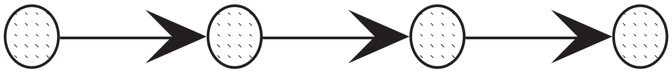
- *Coincidental:*

  Several things happen to describe same object and have no determinative relationship between them

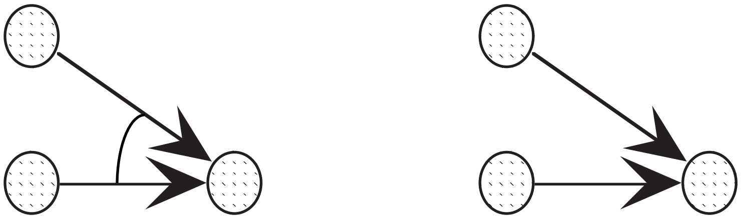
- *Casual:*

  One thing causes another thing to happen

  - *Chaining:*



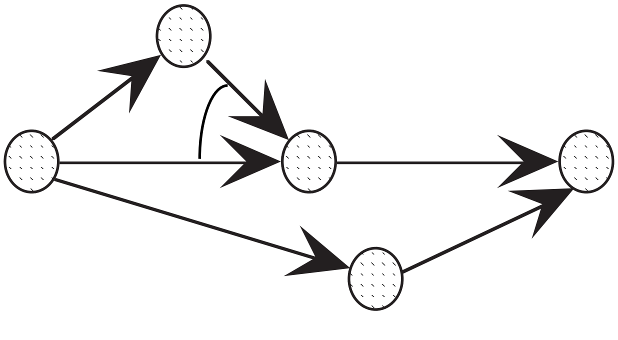
  - *Conjunction (Confluence):*



AND

OR (redundant causation)

  - *Network:* A network of events



- *Preventive:* One thing prevents another

  e.g.; *She prevented the catastrophe*

Recognizing and defining causality is difficult

- *Simultaneous Plant Death:*
  - My rose bushes & my neighbor’s rose bushes both die
  - Did the death of one cause the other to die?

- *Drought:*
  - There has been a drought.
  - My rose bushes and my neighbor’s rose bushes both die.
  - Did the drought cause both rose bushes to die?

- *Traffic:*

  My friend calls me up on the telephone and asks me to drive over and visit. While driving, I ignore a stop sign and drive through an intersection. Another driver hits me. I die.

  Who caused my death? -- Me? -- Other driver? -- My friend? -- Traffic engineer who designed the intersection? -- Fate?

- *Poison:* (Chance increase without causation)
  - Fred and Ted both want Jack dead.
    > Fred poisons Jack’s soup
    > Ted poisons Jack’s coffee
  - Each act increases Jack’s chance of dying
  - Jack eats the soup <u>but</u> leaves the coffee, and dies later
  - Ted’s act raised chance of Jack’s death but was not the cause

- Varying definitions:
  What makes a causal relationship

- Causal asymmetries often play a part

  - *Time order:*
    Effects do not come before effects

  - *Probabilistic independence:*
    Event causes are probabilistically independent while effects are probabilistically dependent

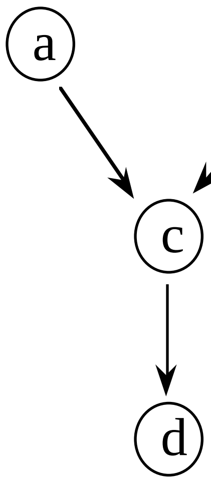
  - *Counterfactual dependency:*
    - Effects counterfactually depend on causes
    - Causes do not counterfactually depend on their effects
    - Effects of a common cause do not counterfactually depend on each other

  - *Over determination:*
    Effects *often* over determine their causes

  - *Connection dependency*
    When breaking connection between:
    cause and *effect*
    only the effect may be affected

Representation: Causal representation underdeveloped

- Representation constrains & supports methods used
- Graphs
  - Some authors suggest that *sometimes* it is possible to recognize causal relations using directed graphs
  - Developing directed acyclic graphs from data is computationally expensive. Work increases geometrically with the number of attributes

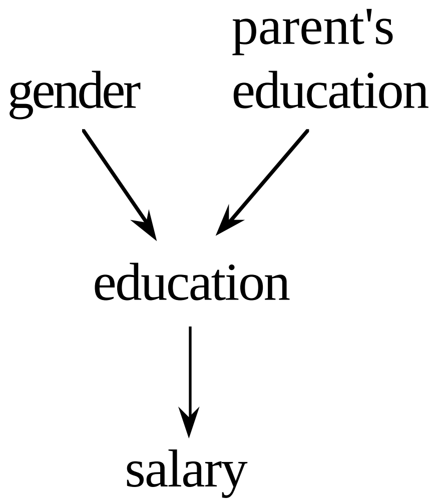


a

b

c

d

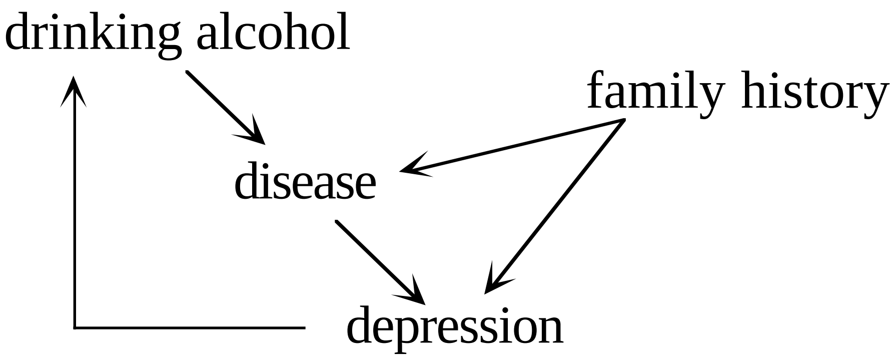


gender

parent's education

education

salary



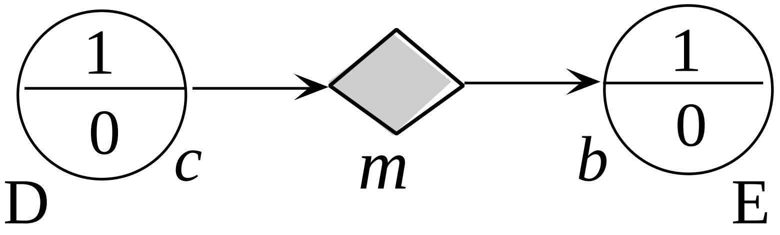
drinking alcohol

family history

disease

depression

- Quantitatively describing relationships between nodes can be complex. Possible general model:

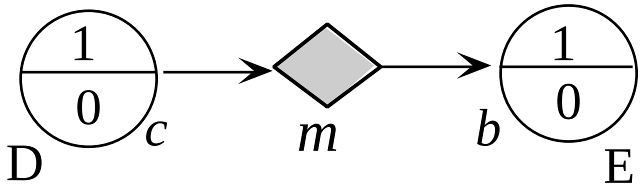


  - *State value 1/0 as event either happens or does not*

  - *State value 1/0 as event either happens or does not*

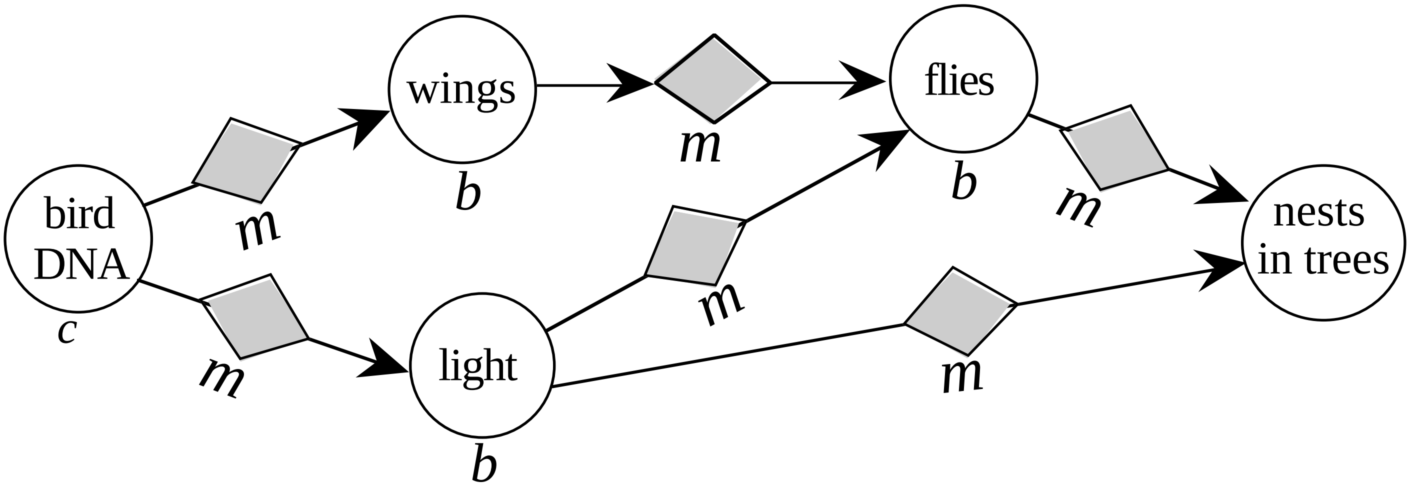
  - *c = P(D)*
  - *m = probability/possibility that when D is present, causal mechanism brings about E*
  - *b = probability that some other (unspecified) causal mechanism brings about E*

- Quantitatively describing relationships between nodes can be complex. Possible general model:



  - *State value 1/0 as event either happens or does not*
  - *c = P(D)*
  - *m = probability/possibility that when D is present, causal mechanism brings about E*
  - *b = probability that some other (unspecified) causal mechanism brings about E*

“bird” example



EPILOGUE

- *Causality* occupies central position in human common-sense reasoning
- Essential role in common-sense human decision-making.
- Deep question:
  - When anything can be said to cause anything else
  - And if it does, what is the nature of causality?
- Whether recognizing causality can be done at all has been a speculation for thousands of years
- At the same time, in our daily lives, we act on common-sense belief that causality exists
- Concern is: How to computationally recognize common-sense causal relationships

- Data mining holds the promise of extracting unsuspected information from data
  - In many ways, interest in association rules is promise (or illusion) of causal, or at least, predictive relationships.
  - Methods build rules
  - Rules indicate co-occurrence strength of data attributes
  - Rules only develop joint occurrence ratios; not causal information
  - *Question:* Whether recognizing an association can lead to recognizing causal relationship
  - *Question:* How to best recognize causality or non-causality in association rules
- *Question:* Determining varying causal strength
  - Strength different in independent relationships
  - Items two way causal relationships

- “**causality**” is like “**truth**” -- words with many meanings and facets.
  - Some of definitions are extremely precise
  - Some involve a reasoning style best supported by approximate reasoning
  - Representations that embrace aspects of imprecision are necessary
- Defining, representing causal & potentially causal relationships necessary to using machine-based methods
- Strong motivation to attempt causality discouvery in mined data

  ⇒High decision value results
- Research concern is how to best approach recognition of causality or non-causality.