Mining Causality From Imperfect Data

LAWRENCE J. MAZLACK
Applied Artificial Intelligence Laboratory
University of Cincinnati
Cincinnati, OH 45221-0030
mazlack@uc.edu

Causal reasoning plays an essential role in both informal and formal human decision-making. Causality itself as well as human understanding of causality is imprecise, sometimes necessarily so. A common sense understanding of the world tells us that we have to deal with imprecision, uncertainty and imperfect knowledge. A difficulty is striking a good balance between precise formalism and commonsense imprecise reality. An algorithmic method of accommodating imprecision in causality is needed. Today, data mining holds the promise of extracting unsuspected information from very large databases. However, the most common data mining rule forms do not express a causal relationship. Without understanding the underlying causality, a naïve use of data mining rules can lead to undesirable actions.

1. Introduction

Causal reasoning occupies a central position in human reasoning. It plays an essential role in human decision-making. Considerable effort has been spent examining causation. Philosophers, mathematicians, computer scientists, cognitive scientists, psychologists, and others have explored causation beginning at least three thousand years ago with the Greeks.

Whether causality can be recognized at all has long been a theoretical speculation of scientists and philosophers. At the same time, in our daily lives, causality is not a speculation; we operate on the commonsense belief that causality exists.

Causal relationships exist in the commonsense world. If an automobile fails to stop at a red light and there is an accident, it can be said that the failure to stop was the accident's cause. However, failing to stop at a red light is not a certain cause of a fatal accident; sometimes no accident of any kind occurs. So, at least some knowledge of some causal effects is imprecise. Perhaps, complete knowledge of all possible factors might lead to a crisp description of whether a causal effect will occur. However, in our commonsense world, it is unlikely that all possible factors can be known. What is more, there appears to be inherent limits on what can be crisply known. None-the-less, we need to deal with common sense cause and effects.

Our common sense understanding has to deal with imprecision, uncertainty and imperfect knowledge. This is also the case for scientific world knowledge. An algorithmic way is needed to deal with causal imprecision. Models accommodating imprecision are needed. These models may be symbolic or graphic. Striking a good balance between precise formalism and commonsense imprecise reality is difficult.

1.1. *Data Mining, Introduction*

Data mining is an advanced tool for managing large masses of data. It analyzes data previously collected. It is *secondary* analysis. Secondary analysis precludes the possibility of experimentally varying the data to identify causal relationships.

There are several different data mining products. The most common are either *conditional rules* or *association rules*. Conditional rules are most often drawn from induced trees while association rules are most often learned from tabular data. Of these, the most common data mining product is association rules; for example:

Customers who
buy *milk*
also tend to buy *bread*

Customers who
buy *strawberries* and *whipped cream*
also tend to buy *cake*

Figure 1. Examples of association rules. Rules can associate as many elements as desired. (Rule quality measures exist but are not shown.)

At first glance, association rules seem to imply a causal or cause-effect relationship. That is: *A customer's purchase of bread causes the customer to also buy milk*. In fact, all that is discovered is the *existence* of a statistical relationship between the items. The *nature* of the relationship is not specified. We do not know whether the presence of an item or sets of items causes the presence of another item or set of items; or the converse, or some other phenomenon causes them to occur together.

When typically developed, association rules do not *necessarily* describe causality. Also, the strength of causal dependency, if any, may be very different from a respective association value. All that can be said is that associations describe the strength of joint co-occurrences. Sometimes, the relationship might be causal; for example, if someone eats salty peanuts and then drinks beer, there is probably a causal relationship. On the other hand, it is unlikely that a crowing rooster causes the sun to rise.

1.2. *Naive Association Rules Can Lead To Bad Decisions*

One of the reasons why association rules are used is to aid in making retail decisions. However, simple association rules may lead to errors. It is common for a food store to put one item on sale and then to raise the price of another item whose purchase is assumed to be associated. This may work if the items are truly associated; but it is problematic if association rules are blindly followed [9].

Example: At a particular 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

[*] Sauerkraut is a form of pickled cabbage. Some people enjoy sauerkraut as a garnish with sausages such as hot dogs. However, it is rarely consumed as a garnish with hamburger. For more about sauerkraut, see: *http://www.sauerkraut.com/*

This would produce the transaction matrix:

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

This would lead to the associations:

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

If the merchant:

- Reduced price of hamburger (as a sale item)
- Raised price of sauerkraut to compensate (as the rule *hamburger, sauerkraut* has a reasonably high confidence.
- The offset pricing compensation would not work as the sales of sauerkraut would not increase with the sales of hamburger. Most likely, the sales of hot dogs (and consequently, sauerkraut) would likely decrease as buyers would substitute hamburger for hot dogs.

2. Causality

Centuries ago, in their quest to unravel the future, mystics aspired to decipher the cries of birds, the patterns of the stars and the garbled utterances of oracles. Kings and generals would offer rewards for the information soothsayers furnished. Today, though predictive methods are different from those of the ancient world, the knowledge that dependency recognition attempts to provide is highly valued. From weather reports to stock market prediction, and from medical prognoses to social forecasting, superior insights about the shape of things to come are prized [3].

Democritus, the Greek philosopher, once said: "Everything existing in the universe is the fruit of chance and necessity." Both randomness and causation are in the world. Democritus used a poppy example. Whether the poppy seed lands on fertile soil or on a barren rock is chance. If it takes root, however, it will grow into a poppy, not a geranium or a Siberian Husky [5].

Beyond computational complexity and holistic knowledge issues, there appear to be inherent limits on whether causality can be determined. For more detail see [7]; briefly, some are:

- Quantum physics
- Impossibility of complete knowledge
- Gödel's theorem
- Turing's halting problem
- Chaos Theory
- Space-Time
- Arithmetic Indeterminism

It may well be that a precise, crisp and complete knowledge of causal events is uncertain. On the other hand, we have a commonsense belief that causal effects exist. If models tolerant of imprecision could be developed, it would be useful. Perhaps, the tools found in soft computing may be useful.

3. Recognizing Causality Basics

A common approach to recognizing causal relationships is by manipulating variables by experimentation. However, non-experimental or observational data is the most likely to be available for data mining analysis. How to accomplish causal discovery in purely observational data is not solved.

Real world events are often affected by a large number of potential factors. For example, with plant growth, many factors such as temperature, chemicals in the soil, types of creatures present, etc., can all affect plant growth. What is unknown is what causal factors will or will not be present in the data; and, how many of the underlying causal relationships can be discouvered among observational data.

An important part of data mining is understanding whether there is a relationship between data items. Sometimes, data items may occur in pairs but may not have a deterministic relationship; for example, a grocery store shopper may buy both *bread* and *milk* at the same time. Most of the time, the *milk* purchase is not caused by the *bread* purchase; nor is the *bread* purchase caused by the *milk* purchase.

Alternatively, if someone buys *strawberries*, this may causally affect the purchase of *whipped cream*. *Some* people who buy *strawberries* want *whipped cream* with them; of these, the desire for the *whipped cream* varies. So, there is a conditional primary effect (*whipped cream* purchase) variably modified by a secondary effect (desire). How to represent all of this is open.

A largely unexplored aspect of mined rules is how to determine when one event causes another. Given that α and β are variables and there appears to be a statistical covariability between α and β, is this covariability a causal relation? More generally, when is any multi-element relationship causal? Differentiation between covariability and causality is difficult.

Some problems with discouvering causality include:

- Adequately defining a causal relation
- Representing possible causal relations
- Computing causal strengths
- Missing attributes that have a causal effect
- Distinguishing between association and causal values
- Inferring causes and effects from a representation.

Complicating causal recognition are the many cases of false causal recognition. For example, a coach may win a game when wearing a particular pair of socks, then always wear the same socks to games.

Beyond data mining, causality is a fundamentally interesting area for workers in intelligent machine based systems. It is an area where interest waxes and wanes; in part because of definitional and complexity difficulties. The decline in computational interest in cognitive science also plays a part. Activities in both philosophy and psychology [2] overlap and illuminate computationally focused work. Often, the work in psychology is more interested in how people *perceive* causality as opposed to whether causality actually exists. Work in psychology and linguistics [4] [6] show that categories are often linked to causal descriptions. For the most part, work in intelligent computer systems has been relatively uninterested in grounding based on human perceptions of categories and causality.

4. Problems With Using Probability

There has been significant work in using various forms of Bayesian networks for causal discovery. A *Bayesian network* is a combination of a probability distribution and a structural model that is a directed acyclic graph where the nodes represent the variables (attributes) and the edges represent probabilistic dependence. In a *causal Bayesian network* the predecessors of a node are interpreted as directly causing the variable associated with a node. However, Bayesian networks can be computationally expensive. Inferring *complete* causal Bayesian networks is essentially impossible in large-scale data mining with thousands of variables.

Restricted algorithms [1] have been suggested that might be useful for causal discovery in market basket data. However, the restrictions on the data and the assumptions made about the relationships are overly limiting. For more detail see [8]; briefly, some are:

- Discrete or continuous data must be reduced to binary values
- There is no missing data
- Causal relationships are not cyclic, either directly or indirectly

5. Epilogue

Causality occupies a central position in human commonsense reasoning. In particular, it plays an essential role in common sense human decision-making by providing a basis for choosing an action that is likely to lead to a desired result. In our daily lives, we make the commonsense observation that causality exists. Carrying this commonsense observation further, the concern is how to computationally recognize a causal relationship.

Data mining holds the promise of extracting unsuspected information from very large databases. Methods have been developed to build rules. In many ways, the interest in rules is that they offer the promise (or illusion) of causal, or at least, predictive relationships. However, the most common form of data mining rules (association) only calculates a joint occurrence frequency, not a causal strength. A fundamental question is determining whether or not recognizing an association can lead to recognizing a causal relationship. A related question is how to determine when causality can be said to be stronger or weaker.

Causality is a central concept in many branches of science and philosophy. In a way, the term "causality" is like "truth" -- a word with many meanings and facets. Some definitions are extremely precise but often unusable for commonsense reasoning. Some involve reasoning style's that may be supported by soft computing.

A deep question is when anything can be said to cause anything else. And if it does, what is the nature of the causality? There is a strong motivation to attempt causality discouvery in association rules. The research concern is how to best approach the recognition of causality or non-causality. Or, if there is any way to recognize causality as long as association rules are the result of secondary analysis?

Acknowledgment

This paper is drawn from work presented at the 2003 FLINT-CIBI Workshop.

References

1. G. Cooper [1997] "A Simple Constraint-Based Algorithm For Efficiently Mining Observational Databases For Causal Relationships," Data Mining and Knowledge Discovery, v 1, n 2, 203-224
2. C. Glymour [2001] **The Mind's Arrows, Bayes Nets And Graphical Causal Models In Psychology**, MIT Press, Cambridge, Massachusetts
3. P. Halpern [2000] **The Pursuit Of Destiny**, Perseus, Cambridge, Massachusetts
4. G. Lakoff [1990] **Women, Fire, And Dangerous Things: What Categories Reveal About The Mind**, University of Chicago Press
5. L. Lederman, D. Teresi [1993] **The God Particle: If the Universe Is the Answer, What Is the Question?** Delta, New York
6. L. Mazlack [1987] "Machine Conceptualization Categories," *Proceedings 1987 IEEE Conference on Systems, Man, and Cybernetics*
7. L. Mazlack [2004] **"**Granular Causality Speculations," *23nd International Conference of the North American Fuzzy Information Processing Society (NAFIPS 2004) Proceedings*
8. L. Mazlack [2004] "Causal Satisficing And Markoff Models In The Context Of Data Mining," *23nd International Conference of the North American Fuzzy Information Processing Society (NAFIPS 2004) Proceedings*
9. C. Silverstein, S. Brin, R. Motwani, J. Ullman [1998] "Scalable Techniques for Mining Causal Structures," *Proceedings 1998 International Conference Very Large Data Bases,* New York, NY, 594-605