Some Issues In Recognizing Causal Relationships

Lawrence Mazlack
Sarah Coppock

Computer Science
University of Cincinnati
Cincinnati, Ohio 45220
{Mazlack,Coppock}@uc.edu

Abstract

Determining causality has been a tantalizing goal throughout human history. Proper sacrifices to the gods were thought to bring rewards; failure to make the proper observations were thought to lead to disaster. Today, data mining holds the promise of extracting unsuspected information from very large databases. The most common build association rules from large data sets. Association rules indicate the strength of association of two or more data attributes. In many ways, the interest in association rules is that they offer the promise (or illusion) of causal, or at least, predictive relationships. However, association rules only calculate a joint occurrence frequency; they do not express a causal relationship. If causal relationships could be discouvered, it would be very useful. Our goal is to explore causality in the data mining context.

Key Words: causality, data mining, association rules, imbalanced causality

1 Introduction

Recognizing when causality occurs implies recognizing a causal relationship; e.g., *A* causes *B* to happen. Whether this can be done at all has been a speculation for thousands of years. At the same time, in our daily lives, we make the common sense observation that causality exists. Carrying this common sense observation further, our concern is how to computationally recognize a causal relationship. Our concern is the discouvery of causal relationships in large data sets.

Causal relationships exist in the common sense world. If someone fails to stop at a red light and there is an automobile accident, we say that the person's failure to stop is the cause of the accident. Another way we think of causal relationships is in a counterfactual sense. For example, if the driver dies in the accident we might say that had the accident not occurred; they would still be alive.

Data mining holds the promise of extracting unsuspected information from very large databases. The most common build association rules from large data sets. Association rules indicate the strength of association of two or more data attributes. In many ways, the interest in association rules is that they offer the promise (or illusion) of causal, or at least, predictive relationships. Whether any association rules express a causal relationship needs to be examined.

Causality occupies a position of centrality in human reasoning. In particular, it plays an essential role in human decision-making by providing a basis for choosing that action that is likely to lead to a desired result.

The ability to form association rules leads to the intuitive desire to infer causal relationships among the items (Mazlack, 2001). A knowledge of causal relationship would be a useful data mining product. However, association rules only describe a joint frequency of the co-occurrence of attribute values. As typically formed, association rules do not represent causal relationships. However, association rules might be useful as a first step in causal discouvery. A discouvered association might flag a potentially interesting relationship; then a causal discovery method might test the relationship.

1.1 Controlled Data

A common approach to recognizing causal relationships is by manipulating a variable through experimentation while observing another variable.

Although there are developed techniques to discouver causal relationships among controlled[1]

[1] *Controlled data* means that the actions producing particular values can be reproduced. Typically, this is experimental data.

data, how to do so in purely observational[2] data is not solved. Observational data is the most likely to be available for analysis; especially in potential data mining applications.

Real world causal events are often affected by a large number of potential factors. For example, with the growth of a plant, many factors such as air temperature, chemicals in the soil, types of creatures present, etc., can all affect plant growth. What is unknown is what causal factors will be present in the data; and, how many of the underlying causal relationships can be discouvered among pure observational data. Taking into account that the data was not collected for the sole purpose of determining causality, it is likely that some factors interacting with the variables will not be present in the data.

In the existing algorithms, assumptions are made in order to infer the relationships. The assumptions often concern the nature of the data distribution. In addition, statistical testing is often used and therefore, the problem of adequate sample (e.g., size of the data set) and selection bias exists.

Some problems with discouvering causality include:
- Defining adequately a causal relation,
- Representing possible causal relations, and
- Infer causes and effects from the representation.

1.2 Discouvery In Observational Data

The present algorithms for discouvery in observational data often use correlation and probability independence to find possible causal relationships. For example, if two variables are statistically independent, it can be asserted that they are not causally related.

One of the limiting factors in being able to infer causal relationships in mining is the need to use observed data rather than controlled data. Another difficulty is that the analysis might be post facto; i.e., after the data has been collected without an opportunity participate in the selection of the recorded variables. Of course, control can be exerted over the attributes selected for consideration from amongst the already recorded data.

Existing algorithms make assumptions and lack some characteristics that may be desirable such as strength of the causal relationship. The issues of adequate sample and selection bias exist.

[2] *Observational data* indicates that the data has been produced and no repeated actions can be taken reproduce further data.

Two important approaches to discouvering causality in observational data, are: constraint-based and Bayesian. The current algorithms do not address any strength or likelihood of the causal relationships inferred. For example, it may be discouvered: *a* causes *b* and *c* causes *d*. Assuming no interrelationships between the two sets, can we say that *a*'s relationship to *b* is stronger or weaker than *c*'s relationship to *d*? If we are using the knowledge for prediction, the strength of the relationship would affect the judgment of prediction.

2 Representation

An open and important issue is the representation chosen. Representation is important as it constrains the discouvery method that can be used. Some representations for use in discouvering causality have been proposed. The representations allow for the characterization and inference of causal relationships. Some representations include:
- Digraphs, such as, directed ascyclic graphs (DAGs), (Sprites, 2000) (Pearl, 2000),
- Probability trees (Shafer, 1998), and
- First-order logic (Hobbs, 2000).

2.1 Digraphs

In a digraph, the vertices correspond to the variables and each directed edge from v_1 to v_2 corresponds to a causal influence from v_1 to v_2. This is shown in *Figure 1*, where both *gender* and *parent's education* have a causal influence on *education*. Pearl (2000) and Sprites (2000) use a form of digraphs called DAGs for representing causal relationships. A DAG consists of a set of vertices, *V*, and a set of directed edges between the vertices, $e(v_1,v_2)$ where v_1,v_2 are in *V*. As a representation for causal relationships, the vertices correspond to the variables (or attributes) that are in a data set. Each directed edge from v_1 to v_2 corresponds to a causal influence from v_1 to v_2. *Figure 1(a)* shows an example of a possible digraph representing causal relationships.

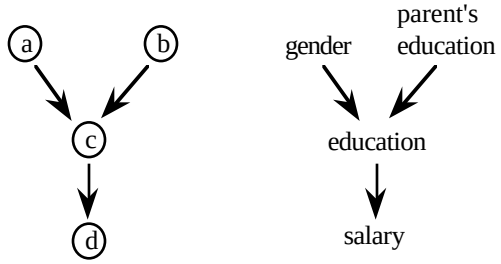

Figure 1(a) An example digraph (DAG)
(b) Example of variables that follow the structure in (a)

Figure 1(b) represents both *gender* and *parent's education* having a causal influence on a person's *education*. According to this representation, *gender* and *parent's education* also have a causal influence

on a person's *salary* even though the relationship is through *education*.

Sometimes, cycles exist. For example, a person's family medical history influences both whether they are depressive and whether they will have some diseases. Drinking alcohol combined with the genetic predisposition to certain diseases influences whether the person has a particular disease, which then influences pain, which thereby may influence the person's drinking habits. *Figure 2* shows the cyclic digraph for this example.

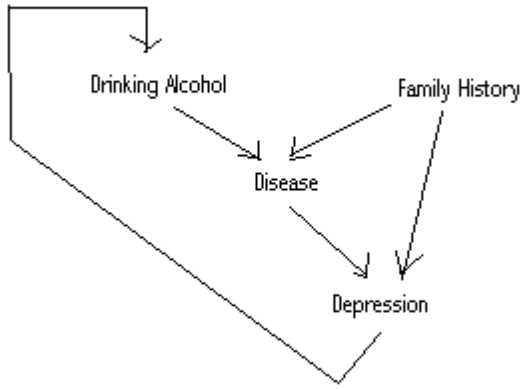

Figure 2. Example of cyclic causal relationships

2.2 Probability Trees

Probability trees can be used to show causal influences Shafer (1998). A tree is a digraph starting from one vertice, the root. In this case, the vertices represent situations. Each edge represents a particular variable with a corresponding probability.

Time ordering of the variables is represented via the levels in the tree. The higher a variable is in the tree, the earlier it is in time. This can become ambiguous for networked representations; i.e., when a node can have more than two parents and thus two competing paths (and their imbedded time sequences).

By evaluating the expectation and probability changes among the situations (or node) in the tree, one can decide whether the two variables are causally related or not.

2.3 First Order Logic

Hobbs (2000) illustrates the use of first-order logic to represent causal relationships. The predicate "causal complex" is defined to indicate a group of events that obtain an effect. This seems a more suitable representation for inference than a DAG or probability tree. One difficulty with this approach is that the representation does not allow for any gray areas. For example, if the event that may have occurred when the wind was blowing east, how could the wind's blowing east-northeast be accounted for? The causality inferred may be incorrect due to the rigidity of the representation. Also, there is no description of the strength of the relationship.

Nor can first order logic deal with dependencies that are only *sometimes* true. For example, *sometimes* when the wind blows *hard*, a tree falls. This kind of *sometimes* event description can possibly be statistically described. Alternatively, a qualitative fuzzy measure might be applied.

Another problem is recognizing differing effect strengths. For example, if it is found that a specific causal complex is discouvered causing a particular event, then is it true that some events in the causal complex are more strongly tied to the effect? Also, it is not clear how a relationship such as the following would be represented: *a* causes *b* some of the time; *b* causes *a* some of the time; other times there is no causal relationship.

3 Nature Of Causal Relationships

Algorithms for discouvering association rules in data with the same observational characteristics have been developed. Since we are able to discouver association rules, the question becomes: can we use this to infer to any degree causal relationships? This would still be valuable to discouver, even if we find that there is no causal relationship.

Some questions about causal relationships that would be desirable to answer are:

- To what degree does *a* cause *b*? Is the value for *b* sensitive to a small change in the value for *a*?
- Does the relationship always hold in time and in every situation? If it does not hold, can the particular situation when it does hold be discouvered?
- Is it possible that there might be mutual dependencies; i.e., $a \rightarrow b$ as well as $b \rightarrow a$? Is it also possible that they do so with different strengths?

When speaking of mutual dependencies, we might describe them as shown in *Figure 3.*

$$a \xrightarrow{S_{a,b}} b \qquad b \xrightarrow{S_{b,a}} a$$

Figure 3. Mutual dependency notation.

where $S_{i,j}$ represents the strength of the causal relationship from *i* to *j* . For example, *a* could be *short men* and *b* could be *tall women*. If $S_{a,b}$ meant the romantic attraction that was caused in *short men* by the sight of *tall women*, we might discouver that

$$S_{a,b} > S_{b,a}$$

It would seem that if there are causal relationships in market basket data, there would often be imbalanced dependencies. For example, if a customer first buys strawberries, there is a reasonably good chance that she will then buy whipped cream. Conversely, if she first buys whipped cream, the subsequent purchase of strawberries is probably less likely.

The present algorithms for discouvery in observational data use correlation between variables and probabilistic independence to find possible causal relationships. If two variables are discouvered to be independent, it is more than likely the two are not causally related. On the other hand, when two variables are found to be correlated, how can it be decided whether one causes the other?

It is potentially interesting to discouver the absence of a causal relationship. For example, assume it is found that shelf position in a food store (e.g., top shelf, bottom shelf, etc.) does not cause a change in the sale of *Cheerios*, but shelf position (e.g. middle shelf) does cause a change in the sale of *Special K*. Then, the decision maker can use this information to adjust their shelves accordingly. A more important use for discouvering lack of a causal relationship is in discouvering causes of disease. If some potential cause can be eliminated, then attention can become more focused on other potentials.

4 Discouvery Based On Graphical Representations

Two approaches to discouvering causality in observational data are *constraint-based* and *Bayesian*. Some algorithms do not address the strength or likelihood of the causal relationships inferred. For example, if it may be discouvered *a* causes *b* and *c* causes *d*, can we say that *a*'s relationship to *b* is stronger or weaker than *c*'s relationship to *d*? If we are using the knowledge for decision-making, the strength of the relationship would affect the decision-making.

4.1 Constraint-Based Discouvery

Using a graphical representation (Pearl, 2000) (Sprites, 2000) (and the underlying algorithms) assumes that the variables satisfy the Markov condition. The Markov condition in this instance means that for every vertice, *v*, given the values of the parents of *v*, *v* is independent of all other vertices *v*' that are not descendants of *v*. In the case of causality, it may be reasonable to assume this. But, this assumption eliminates the potential discouvery of cyclic causal relationships. Because cycles may exist in the real world, they could exist in the data.

It is possible for underlying cycles to exist. For example, a person's family medical history influences both whether they are depressive and whether they will have some diseases. Drinking alcohol combined with the genetic predisposition to certain diseases influences whether the person has a particular disease, which then influences pain, which may therefore influence the person's drinking habits. *Figure 2* shows the digraph for this example. Although cycles may exist in the real world, we cannot assume that they have been captured in the data. If the data set includes all of the variables involved in the cycle, then it will exist. On the other hand, there may be an attribute involved in the cycle missing from the data.

4.2 Reductive Discouvery

A reductive discouvery algorithm contains all possible edges and then selectively eliminates some of them. The PC (Spirtes, 2000) and the IC (Pearl, 2000) algorithms derive a possible underlying DAG by constraining the edges in the graph. For each pair of vertices, the smallest set of vertices on which the two variables become independent is computed. This information is used to either add an edge or remove an edge.

In the PC algorithm, an edge is removed if the two vertices are *d-separated* given a non-empty set of vertices. Two vertices *x* & *y* are said to be *d-separated* given a set of vertices *Z* if *Z* blocks every path from *x* to *y*. *Z* is said to block a path *p* between *x* and *y* if

- *p* contains a chain, $i \rightarrow j \rightarrow k$ or *i* ? *j* ? *k*, or a fork, $i \leftarrow j \rightarrow k$ where *j* is in *Z*; or
- *p* contains an inverted fork, $i \rightarrow j \leftarrow k$, such that *j* is not in *Z* and no descendent of *j* is in *Z*.

Edges are then oriented according to the sets of nodes that separate each pair of nodes. D-separation is a graphical way of expressing conditional independence among variables.

4.3 Additive Discouvery

Additive discovery is the complement of reductive discovery. It starts with nothing and adds DAGs as their need is discouvered

The IC algorithm begins with no edges. If no set of vertices is found that causes an unconditional independence for two vertices, then an edge is added between the two vertices. In other words, the two variables are always dependent with the given variables. After adding all undirected edges as possible, the edges are then oriented according to the sets found in the previous step. In other words, direction is added to the edges according to set of given variables for which the two vertices become independent. The neighboring edges also play a role in which direction the arrow is added between the two vertices.

Both the IC (Pearl, 2000) and the PC (Spirtes, 2000) algorithm have been modified to take into account latent (or unmeasured) variables. Again, this assumption of missing variables is a necessary assumption. The resulting graphs are more complicated. Rather than a resulting partially oriented graph, a partially oriented and partially marked graph results.

4.4 LCD Based Algorithms

The approach of Silverstein (1998) for mining causal structures builds on the LCD algorithm (Cooper, 1997) while retaining its polynomial time complexity. These algorithms do not attempt to discouver the complete causal structure as the IC and PC algorithms do. The LCD algorithm finds causal structures in the form of chains and forks. Without added complexity, an additional causal structure is found; the authors term this structure CCU causality. CCU causality is represented by the structure $a \rightarrow b \leftarrow c$, and is also referred to as *v-structures* (Pearl, 2000).

The authors use support[3]. This restricts the testing of causality to items that would be of more interest. A confidence threshold value is used as the confidence level of the statistic used in determining whether two variables are dependent. The chi-squared (χ^2) statistic is used for determining dependence of two variables. ß² is the statistic computed as: $(O\text{-}E)/E$ where O is the observed value for the variables and E is the expected value for the variables. If ß² is greater than $\chi^2_{\alpha,}$ where á is the confidence level of the test, then the two variables are dependent. Using this statistic in combination with support and confidence, the error of determining dependence is reduced.

Silverstein (1998) doesn't assume that in the case $a \rightarrow b$, a is a direct cause. This allows for possible hidden variables. Other assumptions are made that would be desirable to eliminate. They assume only Boolean data with no missing data. For example, this method could be used on Boolean data extracted from market basket data.

By eliminating the use of graphical representation, the approach is more computationally feasible. But, it does not overcome the difficulty in identifying structure with a certain amount of assurance. Lastly, these methods only assume certain statistical correlation's in a particular orientation among the variables to decide if one variable causes another. Usually, the time ordering of the variables is assumed known beforehand. When the time order is not known, Pearl (2000) introduces the notion of statistical time, any variable ordering that coincides with the causal structure.

[3] *Support* means that particular values must be above a certain frequency in the data.

5 Bayesian Causal Discouvery

A Bayesian approach as described by Heckerman (1995) entails finding the most probable causal structure and the corresponding parameters for the structure. The first complexity issue encountered is the number of possible models. For just 3 variables, there are 25 possible models. Clearly, it would be infeasible to enumerate all of the possible models for even a small data set. To help solve this problem, Heckerman (1995) mentions different approaches have been proposed such as selecting only one "best" model according to the user's knowledge and then computing the parameters. Selecting only a particular portion of the possible models and searching over the selected as though they are exhaustive is another option.

As with many data mining methods, this places an over emphasis on the use of a human expert. Even with an expert, bad choices might be made. Unimportant factors might be included; important factors ignored. A dependence on expert opinion also begs the question. If we are certain that we know the causal factors, there is little reason to run a computationally expensive discouvery algorithm. It is because we are unsure of the causal factors that we need to investigate a solution by using discouvery algorithms.

The Bayesian approach can allow for missing data. However, allowing missing data also increases the complexity issues even larger than previously. One other thing to note, this approach also assumes the nature of variables' distribution in addition to the difficulty of assessing priors to the models. Geiger (1995) shows that there is a known distribution that is suited to be used as priors given independence of the parameters. This distribution is the Dirichlet distribution. Heckerman (1995) shows how to use this distribution to derive the priors for the parameters and to update the priors.

6 Epilogue

Causality occupies a central position in human common sense reasoning. In particular, it plays an essential role in human decision-making by providing a basis for choosing that action that is likely to lead to a desired result.

Recognizing when causality occurs implies recognizing a causal relationship. Whether this can be done at all has been a speculation for thousands of years. At the same time, in our daily lives, we make the common sense observation that causality exists. Carrying this common sense observation further, our concern is how to computationally recognize a causal relationship. Our concern is the discouvery of causal relationships in large data sets.

Today, data mining holds the promise of extracting unsuspected information from very large databases. Methods have been developed to build association rules from large data sets. Association rules indicate the strength of association of two or more data attributes. In many ways, the interest in association rules is that they offer the promise (or illusion) of causal, or at least, predictive relationships. However, association rules only calculate a joint probability; they do not express a causal relationship. If causal relationships could be discouvered, it would be very useful.

Discouvering causal relationships is of great interest. When making decisions, it is often useful to know the relationships between events or items so that this information can be taken into account. If it is found that *a* causes *b*, and *b* is not desired, then this becomes useful information. The interest in defining and discouvering causality is displayed by the amount of literary discussion in fields such as statistics, philosophy, social sciences and others.

Although there are techniques developed to discouver causal relationships among controlled data; how to do so in purely observational data is not solved. Graph based methods are excessively complex when faced by a typical large data set. Our final paper will explore these issues in greater depth.

We are particularly interested in determining when causality can be said to be stronger or weaker. Either in the case where the causal strength may be different in two independent relationships; or, where in the case where two items each have a causal relationship on the other.

To answer these questions, we also must consider how to: recognize when there is a causal relationship, measure the causal strength, and represent the relationship. All of these are open issues on which we are working.

References

R. Agrawal, T. Imielinski, A. Swami [1993] "Mining Association Rules Between Sets Of Items In Large Databases," Proceedings Of ACM SIGMOD Conference On Management Of Data (SIGMOD-93), 207-216

G. Cooper [1997] "A Simple Constraint-Based Algorithm for Efficiently Mining Observational For Causal Relationships" in *Data Mining and Knowledge Discouvery*, v 1, n 2, 203-224

D. Geiger, D. Heckerman [1995] "A Characterization Of The Dirichlet Distribution With Application To Learning Bayesian Networks," in *Proceedings of the 11th Conference on Uncertainty in AI*, Montreal, Quebec, 196-207, August

C. Glymour, G. Cooper, eds. [1999] **Computation, Causation, and Discouvery,** AAAI Press, Menlo Park, California

D. Heckerman [1995] *A Tutorial On Learning With Bayesian Networks*, Microsoft Research Paper, MSR-TR-95-06.

L. Mazlack [2001] "Considering Causality In Data Mining," WSES/IEEE Multi-Conference, Crete, 493-498

J. Pearl, J. [2000] **Causality: Models, Reasoning, And Inference**, Cambridge University Press NY, NY.

G. Shafer [1998] "Mathematical Foundations For Probability And Causality" *Proceedings Of Symposia In Applied Mathematics*, v 55, 207-270

C. Silverstein, S. Brin, et al. [1998] "Scaleable Techniques For Mining Causal Structures," Proceedings 1998 *International Conference Very Large Data Bases*, NY, 594-605

P. Spirtes, C. Glymour, R. Scheines [2000] **Causation, Prediction, And Search**, MIT. Cambridge Massachusetts.