Resumen: On Learning Causal Models From Relational Data - Lee, Honavar

Mauricio Gonzalez Soto

29 de agosto de 2017

1. Previo

- Se estudian Modelos Relacionales Causales (RCM) a través de sus contrapartes relacionales *adjacenty-faithfulness* y *orientation-faithfulness*, lo cual resulta en un enfoque sencillo para identificar un subconjunto de queries relacionales requeridas para determinar la estructura de un RCM utilizando criterios de d-separación sobre un grafo acíclico no-dirigido "desenrrollado" que representa el RCM
- Maier et al (2010) introdujeron RPC, que extiende el modelo PC (Spirtes, 2000) al caso de datos relacionales para aprender relaciones causales.
- RPC utiliza el modelo DAPER (directed acyclic probabilistic entity-relationship) que a su vez extiende el modelo entity-relationship (Chen 1976) para incorporar dependencias probabilisticas
- Maier et al (2013) demostraron que RPC no es completo para aprender modelos causales a partir de datos relacionales, e introdujeron Relational Causal Discovery (RCD) como alternativa
- RCD utiliza un enfoque de restricciones
- Maier et al introdujeron d-separación relacional, que es la contraparte relacional de la d-separación tradicional sobre grafos
- Se introduce además la Abstract Ground Graph
- La prueba del buen funcionamiento de RCD requiere que la AGG represente exactamente todos los vértices que podrían aparecer en todas las instancias el RCM
- Resulta que existen casos en los cuales la d-separación en la AGG no implica las relaciones de independencia condicional que ocurren en el RCM

2. Notación

- Un esquema relacional $\mathcal{S}$ es una tupla $(\mathbf{E}, \mathbf{R}, \mathbf{A}, c)$ formada por un conjunto de clases de entidad $\mathbf{E}$, un conjunto de clases de relaciones $\mathbf{R}$, atributos $\mathbf{A}$ y una función $c : R \times E \to \{una, muchas\}$.
- Denotamos como $\mathbf{I} = \mathbf{E} \cup \mathbf{R}$ y como I_X un objeto I con atributo X
- Un esqueleto relacional $\sigma \in \Sigma_{\mathcal{S}}$ es una instanciación del esquema relacional $\mathcal{S}$ y está representado por un grafo de entidades y relaciones, donde $\Sigma_{\mathcal{S}}$ representa todos las posibles instancias de $\mathcal{S}$.
- Denotamos como $\sigma(I)$ un conjunto de elementos en σ de clase $I \in \mathbb{I}$.
- Un modelo relacional causal (RCM) denotado como $\mathcal{M}$ consiste en un conjunto de relaciones causales $\mathbf{D}$ en el cual causas y efectos están relacionados dado un esquema relacional $\mathcal{S}$

- Un camino relacional P es una sucesión en la cual se alternan clase de entidad y clase de realcion; $P = [I_j, ..., I_k]$. A I_j se le conoce como clase base, y a I_k como clase terminal. Un camino relacional corresponde a una caminata sobre el esquema y muestra cómo la clase terminal está relacionada a la clase base.
- Una variable relacional $P.X$ consiste en un camino relacional P y una clase de atributos X de la clase terminal de P
- Se dice que una variable relacional es canónica si la longitud de su camino relacional es 1.
- Una dependencia relacional especifica causa y efecto; es de la forma $[I_j, ..., I_k].Y \to [I_j].X$; es decir, causa y efecto comparten la misma clase base, y el efecto es canónico.
- Ejemplo, "el éxito de un producto depende de las habilidades de los empleados que lo desarrollan" se representa como

$$[Producto, Desarrollan, Empleados].Habilidades \to [Producto].Exito$$

- Se dice que un RCM es acíclico si existe un órden parcial sobre las clases de atributos **A**. Un RCM aciclico no permite dependencias que conectan una clase de atributos con ella misma
- Un ground graph es una instancia del RCM dado un esqueleto; es decir, se obtiene de interpretar las causas de las dependencias en el RCM sobre el esqueleto al utilizar los conjuntos terminales de cada uno de los elementos del esqueleto.
- Dado un esqueleto relacional σ, el conjunto terminal de un camino relacional P dado un elemnto base $b \in \sigma(P)$, que se denota $P|_b$ está formado por los elementos alcanzables a partir de b cuando se recorre el esqueleto a lo largo de P sin tocar un elemento dos o más veces.
- Denotamos como $\mathcal{GG}_{\mathcal{M}_\sigma}$ la ground graph de un RCM $\mathcal{M}$
- En $\mathcal{GG}_{\mathcal{M}_\sigma}$ existe un vértice $i_j.X \to i_j.Y$ si y sólo si existe una dependencia $[I_k, ..., I_j].X \to [I_k].Y$ tal que $i_j \in \sigma(I_j)$ es alcanzable desde $i_k \in \sigma(I_k)$ a lo largo del camino relacional $[I_k, ..., I_j]$

3. Independencia Condicional en un RCM

- Un RCM $\mathcal{M}$ puede entenderse como un modelo meta-causal definido sobre un esquema $\mathcal{S}$. Dado un esqueleto σ del esquema, el RCM es instanciado en un ground graph, que corresponde a un grafo acíclico no dirigido
- Dados los atributos y valores de los elementos y la estructura del esqueleto, las relaciones de independencia condicional que se cumplen en los datos son equivalentes a aquellas relaciones que cumplen la d-separación sobre $\mathcal{GG}_{\mathcal{M}_\sigma}$
- La contraparte relacional de la d-separación en $\mathcal{M}$ se reduce a la d-separación tradicional sobre todas las instancias de la siguiente manera:
- Si $\mathbf{U}, \mathbf{V}, \mathbf{W}$ son conjuntos disjuntos de variables relacionales. Entonces, para un modelo relacional $\mathcal{M}$, $\mathbf{U}$ y $\mathbf{V}$ están relacionalmente d-separadas por $\mathbf{W}$ si y sólo si para cada esqueleto σ, $U|_b$ y $V|_b$ están d-separados por $W|_b$ en el ground graph para todo $b \in \sigma(B)$.

4. Grafo desenrrollado de un RCM

- Dado un RCM $\mathcal{M} = (\mathcal{S}, \mathbf{D})$, su grafo desenrrollado es un DAG denotado $\mathcal{G}_\mathcal{M}$ en el cual los vértices son variables relacionales de $\mathcal{S}$ y existe un vértice $P.X \to Q.Y$ si y sólo si existe una dependencia $R.X \to [I_Y].Y \in \mathbf{D}$

5. Fidelidad de la representación

- Decimos que una distribución de probabilidad es bien representada (faithfully) por un DAG si las independencias condicionales están bien representadas por el grafo.
- Resulta que un RCM no es fiel a sus AGG, pero se cumplen dos nociones más débiles: fidelidad de adjacencia y de orientación
- Lema 2: si U, V son dos variables relacionales con la misma clase base y son adyacentes en $\mathcal{G}_{\mathcal{M}}$, entonces son condicionalmente dependientes en cualquier subconjunto de $\mathcal{V}_B - \{U, V\}$, donde $\mathcal{V}_B$ es el conjunto de todas las variables relacionales con clase base B
- Lema: Si U, V, W son variables relacionales distintas (con misma clase base B) y se cumple (en $\mathcal{G}$) que U y W están conectadas a V, pero no entre sí (unshielded triple), entonces: si $U \to V \leftarrow W$ (unshielded collider), entonces U y W son dependientes dado cualquier subconjunto de variables relacionales que contenga a V; o, en caso contrario U y W son dependientes.

6. Aprendiendo un RCM

La idea es desarrollar un algoritmo que identifique dependencias no-dirigidas y las oriente utilizando los unshielded collider en el DAG.

- Sea $D = P.X \to [I_Y].Y$, denotamos al reverso de P como $\tilde{P}$ y $\tilde{D} = \tilde{P}.Y \to [I_X].X$
- Decimos que una dependencia es no-dirigida si D y $\tilde{D}$ son válidas
- Se dice que un grafo es parcialmente acíclico dirigido si no existen ciclos dirigidos, pero los vértices pueden ser dirigidos o no.
- Denotamos $X \prec Y$ si existe un camino dirigido de X a Y
- Proposición 1: Sean $[B].X$ y $Q.Y$ variables relacionales distintas con misma clase base B tales que Y no es descendiente de X (en el orden parcial inducido por $\mathcal{M}$). Entonces, $[B].X$ y $Q.Y$ están relacionalmente d-separados por $Pa([B].X)$ y sólo si $Q.Y \to [B].X$ o $\tilde{Q}.X \to [I_Y].Y$ no pertenece a $\mathcal{M}$.

6.1. Fase 1: Identificar dependencias no-dirigidas

Los algoritmos basados en independencias condicionales para aprender modelos causales empiezan enumerando todos los posibles dependencias candidato (Spirtes, Glymour, Scheines 2000). A diferencia el caso proposicional, en el cual el número de variables es fijo y finito, en el caso relacional las variables relacionales son infinitas, por lo que es imposible enumerarlas. (Maier et al. 2013) asume que el numero dependencias en un RCM es finito, y que la longitud del camino más largo en el grafo es conocido.

- Lema: Sea $D = P.X \to [I_Y].Y$. Entonces, $P.X$ es condicionalmente independiente de $[I_Y].Y$ dados los padres de $[I_Y].Y$ en $\mathcal{M}$ o $\tilde{P}.Y$ es condicionalmente independiente de $[I_X].X$ dados los padres de $[I_X].X$ si y sólo si D y $\tilde{D}$ no están al mismo tiempo en $\mathbb{D}$.

6.2. Fase 2: Orientar las dependencias utilizando independencias condicionales

Sea $\mathcal{G}_{\tilde{\mathcal{M}}}$ el grafo desenrrollado de $\mathcal{M} = (\mathcal{S}, \hat{D})$ con $\hat{\mathbf{D}}$ que se obtiene de la Fase 1. Vamos a utilizar el Lema 2 para orientar las dependencias no dirigidas. El siguiente Lema muestra cómo detectar colliders.

- Sean (U, V, W) un "unshielded triple" en $\mathcal{G}_{\tilde{\mathcal{M}}}$. Si existe un conjunto separador tal que U y W sean condicionalmente independientes dado S en $\mathcal{M}$ y V no pertenece a $\mathbf{S}$, entonces $U \to V \leftarrow W$ en $\mathcal{G}_{\mathcal{M}}$.

Resulta que $\mathcal{G}_{\hat{\mathcal{M}}}$ es un grafo infinito, por lo que no podemos aplicar directamente detección de colisionadores. Pero resulta que para cada tripleta unshielded existe una tripleta representativa tal que orientarla sea equivalente a orientar las tripletas en $\mathcal{G}_{\hat{\mathcal{M}}}$

- Lema 5: Si $(P'.X, Q'.Y, R'.Z)$ es una tripleta unshielded en $\mathcal{G}_{\hat{\mathcal{M}}}$, existe una tripleta representativa $([I_X].X, Q.Y, R.Z)$ en $\mathcal{G}_{\mathcal{M}}$.

La existencia de estas tripletas representativas permite darle orientación a las dependencias relacionales en los colisionadores unshielded del RCM sin la necesidad de buscar estas tripletas unshielded sobre las AGG.

Ahora, combinando el Lema 5 y la Proposición 1 tenemos que

- Corolario 1: Sea $([I_X].X, Q.Y, R.Z)$ una tripleta unshielded en $\mathcal{G}_{\hat{\mathcal{M}}}$ tal que X no precede a Z en el orden parcial del grafo. Entonces, $([I_X].X)$ es condiciomalmente independiente de $R.Z$ dados los padres de $[I_X].X$

Como la existencia de un colisionador unshielded $([I_X].X, Q.Y, R.Z)$ imploca la existencia de otro, $([I_Z].Z, \tilde{D}_2.Y, \tilde{R}.X)$ donde $D_2.Z \to [I_Y].Y$ está en $\hat{\mathbb{D}}$, entonces uno puede darle orientación a las dependencias entre X y Y y entre X y Z sin importar el orden entre X y Z

6.3. Fase 3

La Fase 2 no sólo orienta dependencias que forman colisionadores unshielded, sino que además impone restricciones sobre pares de dependencias que forman tripletas unshielded. Lo que se hace ahora es traducir la información expresada utilizando variables relacionales a información descrita utilizando los atributos. Primero, representamos no-colisionadores unshielded como colisionadores unshielded sobre atributos.

Se introduce una *class dependency graph* $\mathcal{G}_\pi$ sobre el conjunto de clases de atributos $\mathbf{A}$ en la cual existe un vértice que une X con Y si existe una dependencia entre X y Y

7. Algoritmo

- **Input:** Un esquema relacional $\mathcal{S}$, una distribución p, un tamaño máximo de pasos h.
- **Salida** un modelo relacional causal parcialmente dirigido
- 1 Inicializar D con las dependencias candidato
- $d = 0$
- Repetir
  - For $D = U \to V \in \mathbf{D}$ do
    - Si U y V son condicionalmente independientes dado $S \subset Pa(V) - \{U\}$ entonces remover $\{D, \tilde{D}\}$ de $\mathbf{D}$
  - $d = d + 1$
- Hasta $|Pa(V)| < d$
- $\mathcal{N} = \emptyset$, inicializar $\mathcal{G}_\pi$ a partir de $\mathbf{D}$ y $\mathcal{K}$
- Aplicar reglas de orientación den $\mathcal{G}_\pi$
- Para tripletas representativas $([I_X].X, Q.Y, R.Z)$ hacer:
  - Si $X \prec Z$, continuar

- Si $X - Y$ y $Y - Z$ están orientadas, continuar
- Si $(X, Y, Z) \in \mathcal{N}$ o $X \leftarrow Y$ o $Y \rightarrow Z$, continuar
- Si $[I_X].X$ ind cond dado $S \subseteq Pa([I_X].X$ entonces
  - Si $Q.Y \notin S$, entonces orientar $X \rightarrow Y$, $Z \rightarrow Y$
  - Si $X = Z$, orientar $Y \rightarrow X$
  - En caso contrario, añadir (X, Y, X) a $\mathcal{N}$
- $X \rightarrow Z$
- Aplicar reglas de orientación

- Orientar **D**