Causal Structure Learning via Temporal Markov Networks

Aubrey Barnard, David Page

University of Wisconsin-Modison

Problemática

La motivación se encuentra en identificar los eventos adversos de drogas (ADE). La Asociación de Resultados Médico Observacionales (OMOP) provee los registros médicos electrónicos (EMR). Ya existen contribuciones a esta tarea, como aprender la estructura DBN causal.

Los desafíos del descubrimiento causal implican: 1) los datos están desordenados en los eventos. 2) los pacientes interactúan con el sistema esporádicamente y solo cuando están enfermos. 3) de las miles de variables que posee EMR solo se registran unas pocas. 4) debido a lo anterior hay ruido.

Contribución

Evitar combinatorias de los algoritmos de búsqueda (estructuras) reformulando la estructura aprendida del problema, es optimización no combinatoria, suave y convexo en un modelo log-lineal.
El aprendizaje de la estructura causal a través de TMNs aborda estos problemas:
(1) Al aprender la estructura dirigida utilizando un modelo no dirigido (TMN) y el aprendizaje de los parámetros es un problema de optimización convexo.
(2) Usar características para modelar la irregularidad, la escasez, y temporalidad de los datos EMR.
La combinación del aprendizaje de estructuras mediante el aprendizaje de parámetros y el modelado temporal aproximado es novedosa

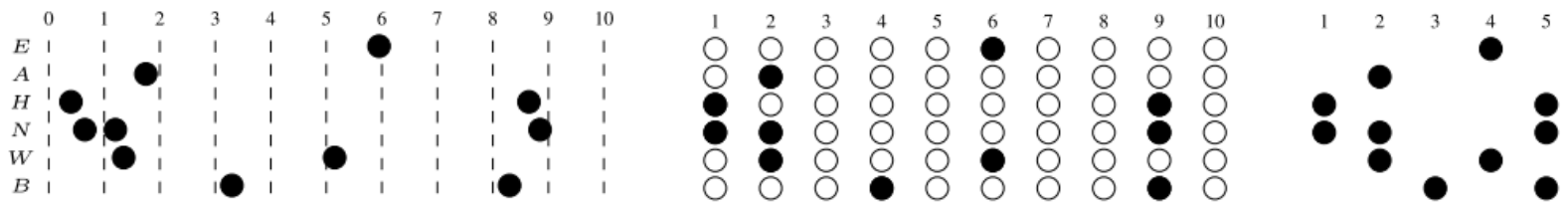
Definición 1 (Propiedad de factorización): Una distribución P(X) se factoriza de acuerdo con un gráfico G no dirigido, si su densidad se puede expresar como un producto de funciones potenciales no negativas sobre las cliques C de G.

$$P(X) \propto \prod_{c \in C} \psi_c(X_c) \tag{1}$$

Definición 2 (Propiedad global de Markov): X⊥Y|Z si Z separa a X y Y en G.
Teorema 3: Para cualquier gráfico G no dirigido y cualquier distribución de probabilidad P en X, se sostiene que la propiedad de factorización implica la propiedad global de Markov.

Los TMN es un tipo log-lineal de Modelo Gráfico Probabilista con funciones características para modelar lineas de tiempos.

Timeline.
Es un conjunto de variables X que ocurren en un tiempo T, esto es S=$X_{i,t}$: (X_i,t)$\in$ XxT.



(a) Point events in continuous time (EMR) (b) Fully-observed discrete time (DBN) (c) Condensed

Figure 2: Various forms of a sequence of events (timeline) as might be observed from a process like Figure 1b.

Modelo log-lineal

Definición 4 (Temporal Markov Network): una TMN es una tupla (X,F,θ) donde X es un conjunto de tipos de eventos (variables aleatorias), F es un conjunto de funciones de funciones binarias $f_i(\mathrm{X}_i \subseteq \mathrm{X}):\mathrm{X}_i \mapsto 0, 1$ y $\theta \in \mathrm{R}^{|F|}$. Se establece en el modelo log-linear en las ecuaciones 2 y 3.

$$P(S = s) = \frac{1}{Z} exp(\sum_i \theta_i f_i(s)) \tag{2}$$

$$Z = \sum_{s \in S} exp(\sum_i \theta_i f_i(s)) \tag{3}$$

$f_i \in$ F, $\theta_i \in \theta$ y X_i es un subconjunto de clique inducido en G.

Funciones Características

Las siguientes características modelan los aspectos más destacados de las líneas de tiempo como predicados lógicos. En la notación, S es una línea de tiempo, T es un paso de tiempo, X, Y, Z son eventos, mayúsculas indican variables y minúsculas indican valores instanciados.

- evento, $f_s(x)$:true, si ocurre S (atemporal)
- evento@, $f_s(x_t)$:true, si el evento x ocurre en un t en S (atemporal)
- co-ocurrencia, $f_s(x, y)$:true, si los eventos x y y ocurren en S (atemporal)
- co-ocurrencia@, $f_s(x_{t1}, y_{t2})$:true, si x ocurre en un tiempo t_1 y y en un tiempo t_2 en S (temporal)
- before, $f_s(x \rightarrow y)$:true, si x y y ocurren en S, pero x antes que y (temporal)

- before-δ, $f_s(x_T, y_{T+\delta})$: true, si x y y ocurren en S y x ocurre δ pasos antes que y (temporal)
- before3, $f_s(x, y \to z)$:true, si x, y y z ocurren en S, y x y y ocurren antes que z (temporal)

$f_s(w, b)$ y $f_s(b, w)$ son redundantes, pero $f_s(b \to w)$ y $f_s(w \to b)$ son ordenados, ambos pueden ser verdad en la figura 1.

Las características @ están ancladas a pasos de tiempo específicos, pero las otras características flotantes siendo menos específico que el anclaje, vincula los parámetros a través de los pasos de tiempo.

Dependiendo de la elección de las características y el parámetro de vinculación que inducen, las TMN pueden representar análogos no dirigidos de BN, DBN y redes de eventos.

Los parámetros se aprenden utilizando la estimación de maximum likelihood estándar. Encontrar el máximo de probabilidad logarítmica es un problema de optimización continuo y convexo, que se resuelve mediante el ascenso por gradiente. Debido a que el maximum log-likelihood es global, se alcanza cuando el gradiente (Ecuación 4) es cero.

$$\frac{\partial}{\partial \theta_i} \frac{1}{|D|} log\mathcal{L}(\theta; D) = E_D(f_i(s)) - E_\theta(f_i(s)) \tag{4}$$

Para calcular el gradiente primero se calcula E_D (la estadistica esperada de los datos) y se hace una vez. Luego la estadistica esperada del TMN (E_θ) y esto se debe hacer por cada cambio de parametro.

Aprendizaje de la estructura causal a través del aprendizaje de parámetros

La detección de la independencia condicional se realiza construyendo una TMN, aprendiendo los pesos de sus características y comparando esos pesos con cero. Un peso cero indica la ausencia de relación modelada por esa característica. Si dos solo son 0 indica que son condicionalmente independientes. Esta propiedad permite que el aprendizaje de peso en las RGT recupere la estructura de independencia condicional de la DBN generadora.

Aprendizaje de la estructura causal a través del aprendizaje de parámetros

Teorema 5 (Aprendizaje de estructura TMN: Dado un DBN M que genera una verdadera distribución P(S) en las líneas de tiempo, los bordes delanteros del DAG G de M se pueden deducir de los pesos de una TMN ajustados a P(S) usando la maximum likelihood. Específicamente, si los pesos de $f_i(X \to Y)$ y todas las demás funciones que contienen X y Y son cero, entonces $X \to Y$ no es una arista en G:

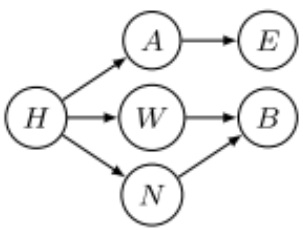
$$(\forall(i : f_i \supseteq X, Y)\theta_i = 0) \Longrightarrow X \to Y \notin G \tag{5}$$

Experimentos

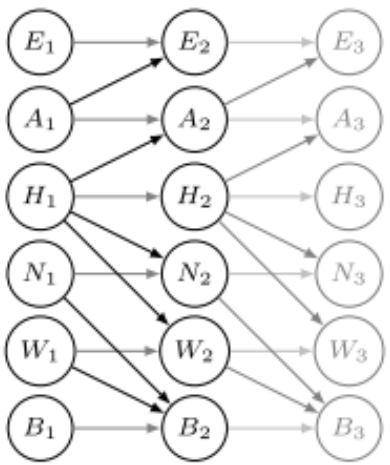
Para evaluar las RGT, se realizaron experimentos para compararlos con otros métodos en tareas de aprendizaje de estructura DBN utilizando datos sintéticos y del mundo real. Se utilizaron para comparar: TMN-PC, TMN-DBN, TMN-Bf3, PC, BNF-DBN (comparar estructuras)

Drug	Condition	Label
A ACE inhibitors	E angioedema	+
T amphotericin B	R acute renal failure	+
I antibiotics	L acute liver failure	+
P antiepileptics	S aplastic anemia	+
Z benzodiazepines	F hip fracture	+
Φ bisphosphonates	U upper GI ulcer	+
D tricyc. antidep.	M acute MI	+
Y typ. antipsycho.	M acute MI	+
W warfarin	B bleeding	+
β beta blockers	X MI mortality	−
N NSAIDs	H hypertension	

(a) OMOP ADE task and list of events (variables)



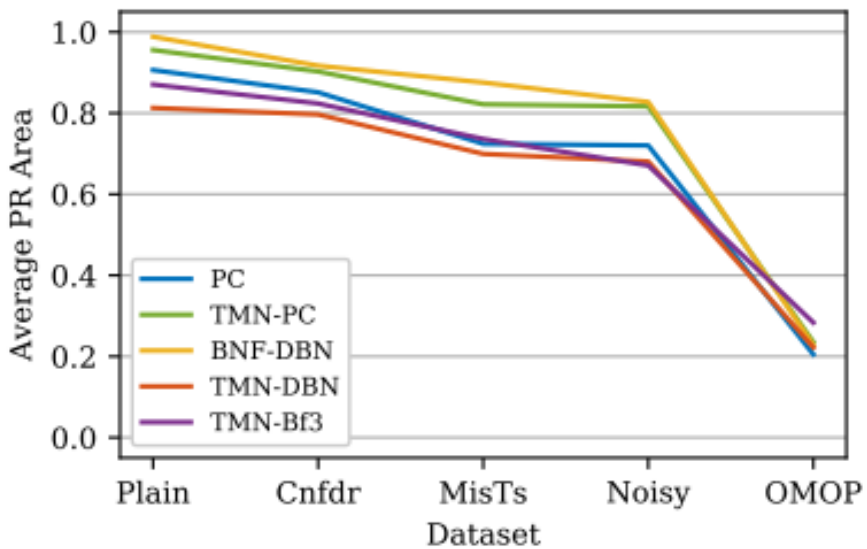
(b) Causal BN



(c) Equivalent DBN

Para evaluar se utilizó precision-recall (PR).

Se realizo una clasificación binaria donde se cada conjunto de pruebas eras realizadas a mano y otras al azar.
Resultados de la prueba con datos sintéticos.



(a) Average PR areas of the methods across data regimes

Name	People	PoI	EoI	Years
GE	11.2M	4.1M	7.1M	1995–2009
CCAE	46.5M	25.6M	47.7M	2003–2009
MDCD	10.8M	7.3M	14.0M	2002–2007
MDCR	4.6M	3.9M	12.7M	2003–2009
MSLR	1.2M	1.1M	2.1M	2003–2008

(b) GE Centricity (EMR), MarketScan Commercial Claims and Encounters (claims), MarketScan Medicaid (claims), MarketScan Medicare (claims), MarketScan Lab (claims). EoI: events of interest. PoI: people with EoI.

Experimento OMOP

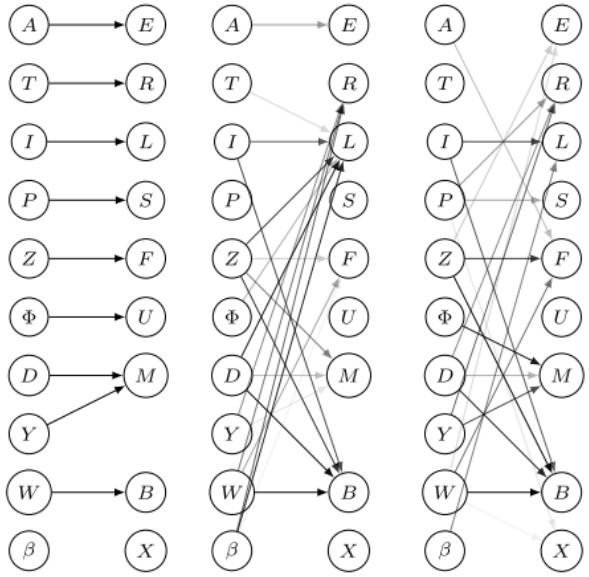
Para crear un conjunto de datos de cada base de datos, se extrajo una línea de tiempo para cada paciente y luego se condensó.

Las variables no observadas se asumieron como falsas. Se extrajeron veinte muestras de 100 mil líneas de tiempo sin reemplazo de cada conjunto de datos.

PC y BNF-DBN no pudieron escalar al tamaño completo de los datos. (Las RGT, que solo necesitan estadísticas suficientes, no tienen limitaciones directas de tamaño de datos).

	X	$\hat{\beta}$	P-Value
1	BNF-DBN? * cnfdr	-0.713	0
2	density, e/n^2	0.663	0
6	TMN-Bf3? * cnfdr	-0.454	3.16e-232
7	BNF-DBN? * noise	-0.414	0
8	BNF-DBN?	0.400	0
9	TMN-PC?	0.309	0
10	BNF-DBN? * mists	-0.305	0
15	PC?	0.245	0
16	TMN-DBN? * noise	-0.225	0
17	TMN-DBN?	0.217	0
19	TMN-Bf3?	0.209	0
21	TMN-Bf3? * mists	-0.157	1.26e-186
22	# cnfdr $/n$	0.130	3.99e-39
23	log # data	0.0747	0
24	missingness	-0.0227	2.16e-09
25	noise level	-0.0213	3.79e-08
32	condensed?	0.000733	0.296

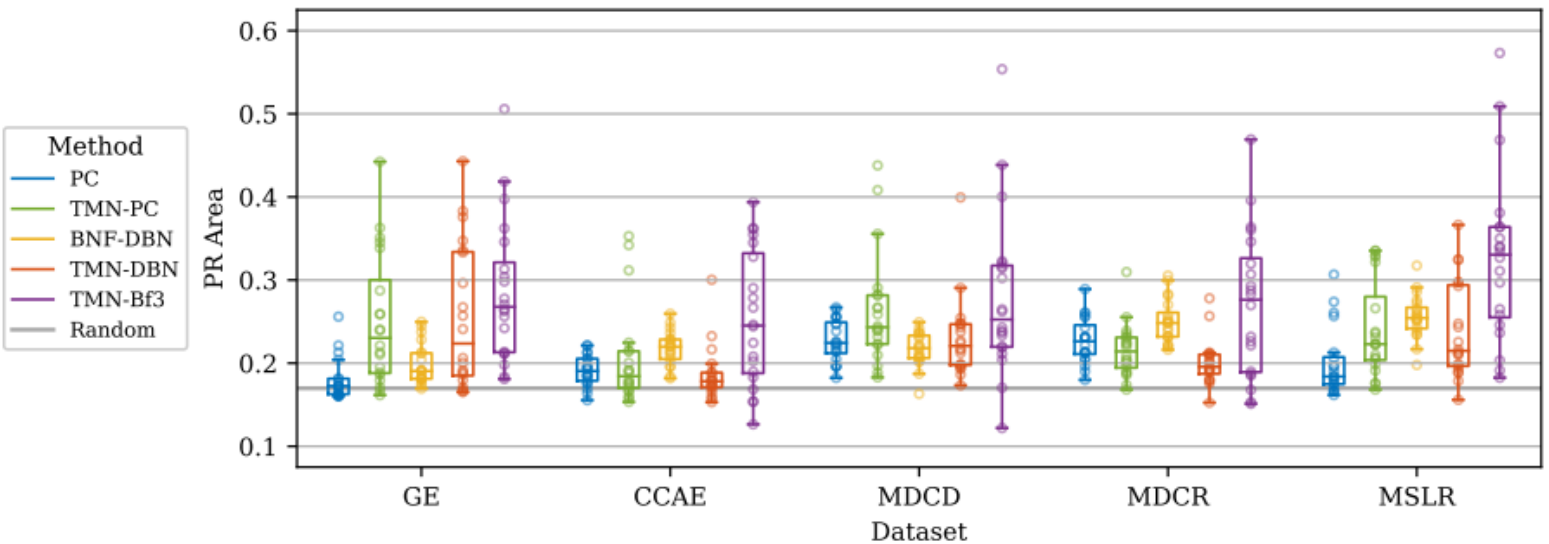
(a) Selected results from a linear regression of PR areas



(b) Ground truth (c) PC, MSLR (d) TMN-Bf3, MSLR

Experimentos: OMOP

Las TMN funcionan especialmente bien con los datos de EMR de GE. Los resultados son variados:
-TMN-PC vence a PC en 3 conjuntos de datos.
-TMN-DBN vence a BNF-DBN en 2 conjuntos de datos.
-TMN-Bf3 es el mejor en los 5 conjuntos de datos.



La hipótesis del éxito de TMN-Bf3 se debe a su capacidad para modelar las interacciones de orden superior y detectar la independencia en presencia de ruido.

Experimentos: OMOP

Los tiempos de ejecución **(min, avg, max)**, en horas, en la tarea OMOP fueron BNF-DBN (0.6, 0.8, 0.9), TMN (0.5, 1.3, 2.8) y PC (0.1, 2.9, 9.9). Si bien esto hace que BNF-DBN parezca rápido, los experimentos tuvieron que limitarse a 100k líneas de tiempo para hacer que BNF-DBN y PC sean manejables.

Discusión

La convexidad garantiza que existe un óptimo global y que no hay impedimentos para llegar allí, como mesetas u óptimos locales. Esto garantiza el progreso con cada iteración, y la optimización se puede detener en cualquier momento para obtener una solución aproximada con el gradiente que da una idea de qué tan cerca está el modelo actual del óptimo.

La formulación como **modelo log-lineal**:
Ventajas:
-Permite características arbitrarias, que pueden usarse para manejar eventos irregulares y modelar dependencias de corto y largo alcance.
-Permite escalar a conjuntos de datos muy grandes al separar el procesamiento de datos de la optimización.
Desventajas:
-Ahora hay un problema de modelado ya que uno debe elegir las características correctas.
-Dependiendo de cuántas funciones se elijan, su complejidad y de cuántas combinaciones de eventos se ejemplifiquen, puede haber una gran cantidad de funciones y un espacio de optimización correspondientemente grande.
-Los desafíos de optimización se amplifican por las dificultades de inferencia de un PGM extremadamente grande y no factorizable.