Survey

Mauricio Gonzalez Soto

7 de diciembre de 2017

Índice

1. Introducción

El problema al cual nos enfrentamos consiste en el aprendizaje de relaciones causales a partir de datos. Los datos que, usualmente, se tienen provienen de fuentes ruidosas por lo que la probabilidad será el lenguaje adecuado para razonar sobre estos datos. Al tratar de averiguar relaciones causales, es común contar con información previa sobre el fenómeno dentro del cual se quiere investigar causalidad, por lo que recurrimos a la estadística bayesiana para el manejo de esta información inicial.

La primera parte de este resumen consiste en Redes Bayesianas, que son objetos matemáticos utilizados para representar distribuciones de probabilidad de manera eficiente; además, permiten incorporar conocimiento inicial sobre las variables aleatorias representadas en la red. Debido a que estas redes son de fácil interpretación, es natural definir modelos causales en términos de estas, pues fácilmente nodos y aristas pueden traducirse en términos de objetos y causas.

Veremos que al intentar obtener relaciones causales a partir de información proveniente del mundo real necesitamos obtener las entidades que se relacionan entre sí de manera causal, por lo que haremos un repaso de técnicas utilizadas en extracción de relaciones.

Por otro lado, en la literatura proveniente de Psicología Computacional y Ciencias Cognitivas, se sabe que los seres humanos obtienen conocimiento causal al encontrarse en situaciones en las cuales deben tomar decisiones de manera secuencial para lograr un objetivo futuro. Se sabe también que el conocimiento causal adquirido es utilizado para mejorar las futuras decisiones tomadas.

Finalmente, se discutirá un poco sobre Aprendizaje por Refuerzo, que es un marco teórico para el aprendizaje automático en el cual un *agente* aprende a realizar una tarea a través de experiencia y de interacción con su entorno. Notando que los modelos causales tradicionales requieren de intervenciones y experimentación en el mundo, y que estas intervenciones son una forma de interacción y alteración del mundo, salta a la vista la relación entre Aprendizaje por Refuerzo y aprendizaje causal. Entonces, resulta natural preguntar cómo pueden aprenderse relaciones causales de manera interactiva de modo que el conocimiento causal adquirido permita a un agente tomar mejores decisiones. La pregunta se convierte entonces en cómo obtener e incorporar conocimiento causal en un problema de aprendizaje por refuerzo. Además, uno se pregunta si es posible utilizar este conocimiento causal obtenido al interactuar sobre el mundo para realizar otro tipo de inferencias causales en ciertos dominios de interés.

2. Causalidad

2.1. Intro

Saber si es posible hablar de relaciones causa-efecto es un tema polémico; los métodos probabilistas tradicionales sólo codifican asociaciones estadísticas, pero la idea La diferencia entre modelos probabilistas y modelos causales se da cuando es de interés tomar acciones que modifiquen el entorno y no nada más observar los valores de las variables.
Según **?**, las dos preguntas fundamentales en Causalidad son:

- ¿Qué evidencia empírica es necesaria para hacer inferencias válidas sobre relaciones causa-efecto?
- Dado que estamos dispuestos a aceptar información causal sobre un fenómeno, ¿qué inferencias podemos obtener a partir de esta información, y cómo?

En cuanto al segundo punto, Pearl considera que hay tres tipos de preguntas a considerar

- Predicciones: si sabemos que una persona ingirió alimentos de dudosa procedencia, ¿se enfermará?
- Intervenciones: si le damos cierto tratamiento a un paciente, ¿mejorará?

- Contrafactuales: ¿el paciente del punto uno se habría enfermado de todos modos si no hubiera consumido esos alimentos?

Pearl argumenta que aunque causalidad suele entenderse como afirmaciones necesarias, estas son hechas en contexto de incertidumbre, por lo que cualquier teoría causal debe estar en términos de probabilidades.

2.2. Un poco de historia

Aristóteles definió 4 *causas* para una cosa, una de ellas llamada *causa eficiente* que estaba definida como aquello que hace la cosa.

En 1690, el filósofo inglés John Locke propuso como definición de causa "aquello que produce alguna idea simple o compleja", y lo que es producido lo definió como efecto.

? resalta la importante diferencia en que Aristóteles hizo énfasis en las causas de una cosa, pero no en el efecto de estas.

Fue David Hume quien se considera que hiz una importante contribución, y él definió la causalidad como una relación entre experiencias más que una entre hechos. Argumentó que no es verificable empíricamente que la causa produzca el efecto, sino sólo que el evento llamado *causa* es invariablemente seguido por aquella experiencia llamada *efecto*.

Hume definió 3 criterios básicos para causalidad:

- Contiguidad espacio-temporal: tanto causa y efecto tienen lugar en una entidad común (espacio y tiempo).
- Sucesión temporal: La causa tiene que preceder temporalmente al efecto.
- Conjunción constante: Causa y efecto ocurren juntos, o no ocurren juntos.

John Stuart Mill argumentó tajantemente que observación sin experimentos no puede probar causalidad. Mill identificó cuatro métodos generales para la identificación de relaciones causales.

2.3. Modelo de Rubin

El modelo de Rubin parte de una población de *unidades* U. En general, estas unidades son los objetos de estudio sobre las cuales actuan las causas. Para hacer más sencilla la descripción del modelo, supondremos que sólo existen dos causas (o niveles de tratamiento) denotadas por t y c. Sea S una variable indicadora de la causa a la cual cada unidad $u \in U$ fue expuesta. Además, consideramos dos variables respuesta Y_t, Y_c. $Y_t(u)$ reoresenta el valor de la respuesta si la unidad u es expuesta a t y análogo para c.

Definición 1. *Se define el efecto de la causa t sobre u relativo a la causa c como*

$$Y_t(u) - Y_c(u).$$

Esta expresión se lee de la siguiente manera: t causa el efecto $Y_t(u) - Y_c(u)$ en la unidad u (relativo al control c).

Notemos que es imposible observar simultáneamente el efecto de t y de c sobre una misma unidad u y a esto **?** le llama el Problema Fundamental de la Inferencia Causal. El problema, según explica Holland, radica en la palabra *observar* pues no podemos saber simultáneamente el efecto de exponer a un enfermo en particular a un nuevo tratamiento y al mismo tiempo el efecto de no exponerlo.

2.4. Modelo de Suppes

En su libro "Probabilistic Theory of Causality", (**?**), Patrick Suppes intenta mejorar el análisis de Hume sobre causalidad; en particular, la condición de conjunción constante. Para Suppes, no importa qué sean las causas y los efectos, sino que se puedan expresar como eventos que ocurren bajo ciertas condiciones. Suppes supone que todas las causas necesariamente preceden en el tiempo a sus efectos. Define una causa *prima facie* de un evento como un evento que precede temporalemnte a este y que está asociado de manera positiva con él. Define como una causa espuria a una causa prima facie tal que es condicionalmente independiente del efecto dada otra causa la cual es anterior en el tiempo a la causa prima facie y que está asociada con el evento dada la causa. Una causa genuina es una causa prima facie que no es espuria.

Definición 2. *Si $r < s$ representan valores de tiempo, se dice que el evento C_r es una causa prima facie del evento E_s si*

$$P(E_s|C_r) > P(E_s).$$

Definición 3. *Un evento C_r se dice que es una causa espuria de un eventi E_s si C_r es una causa prima facie de él y para alguna $q < r < s$ existe un evento D_q tal que*

$$P(E_s|C_r, D_q) = P(E_s|D_q)$$

y además,

$$P(E_s|C_r, D_q) \geq P(E_s|C_r).$$

Por lo tanto,

Definición 4. *Para tiempos $r < s$, un evento C_r se dice que es una causa genuina de un evento E_s si C_r es una causa prima facie de E_s, pero no es una causa espuria.*

Notemos que Suppes está definiendo la causa de un efecto, más que el efecto de una causa. Además, no hay restricciones a la naturaleza de una causa más allá de que sea un evento que ocurre antes en el tiempo. El modelo de Suppes no puede expresar el efecto de una causa en particular, sólo comportamientos promedio (**?**).

En palabras de Holland, la causalidad de Suppes no es más que correlación entre causa y efecto que no puede ser eliminada al parcializar causas legítimas.

2.5. Modelos Causales en Medicina

Un caso de aplicación en medicina es determinar que ciertas bacterias causan ciertas enfermedades. Existe un conjunto de postulados planteados por Koch para decidir cuándo un micro-organismo está implicado en una enfermedad. Aunque no existe un planteamiento formal, estos pueden expresarse como (**?**):

- El organismo debe estar presente en todos los casos de la enfermedad supuestamente provocada por él.
- El organismo debe aislarse de los pacientes y cultivarse en estado puro.
- El organismo cultivado en estado puro debe reproducir el padecimiento.

Estos postulados aseguran que no haya datos que apoyen una hipótesis nula de cero efecto causal.

Por otro lado, Hill, quien fue de los primeros en estudiar la asociación entre fumar y cáncer, encontró 9 factores útiles a la hora de decidir que una asociación observada es causal.

2.6. Causalidad de Granger en Economía

La causalidad de Granger se utiliza para series de tiempo, lo cual es de alta importancia en Economía. **?** opina que las ideas esenciales de Granger no están limitadas a series de tiempo y que pueden ser expresadas en términos de variables que actuan sobre unidades de una población.

En la teoría de Granger, una variable causa otra variable; es decir (**?**), las realizaciones de una variable mejoran las predicciones hechas sobre la otra variable. La única parte en la que se utilizan series de tiempo en esta definición es para hablar de valores de las variables definidos hasta un cierto tiempo t.

Si X, Y, Z son variables aleatorias definidas sobre cierta población, se dice que X no es una causa de Granger de Y si X y Y son condicionalmente independientes dado Z. Entonces, X es una causa de Granger de Y si distintos valores de X llevan a distintas distribuciones predictivas de Y dados X y Z; esto es, X ayuda a predecir Y aun cuando Z es tomada en cuenta. La no-causalidad de Granger es parecida a las causas espurias de Suppes pues ambas consideran que no se pueda predecir un evento dada cierta información.

En opinión de Holland, la causalidad de Granger falla de la misma manera que la de Suppes, pues a la luz de nueva información una causa Granger puede dejar de serlo.

2.7. Modelos Causales en Ciencias Sociales

2.8. Aprendizaje de Modelos Causales a partir de datos

En los métodos de estado del arte de Modelos Causales Funcionales suele hacerse los siguientes supuestos (**?**):

- Suficiencia Causal (CSA)
- Propiedad Markoviana (CMA)
- Independencias Condicionales
- Fiabilidad causal (CFA)

2.8.1. Aprender el CPDAG

Existen tres familias de métodos para extraer el grafo: métodos de restricciones, métodos de score y métodos híbridos.

Los métodos basados en restricciones explotan las independencias condicionaes para identificar todas las v-estructuras. El algoritmo más conocido es el Algoritmo PC (**?**). El algoritmo PC primero construye un esqueleto basado en independencias condicionales; luego, identifica las v-estructuras y posteriormente orienta las aristas. El algoritmo FCI (**?**) extiende el algoritmo PC al relajar la suficiencia causal e incorporar variables latentes. El algoritmo RFCI de **?** es más rápido que FCI y maneja datos de alta dimensionalidad. La debilidad de estos métodos, según **?** es que dependen demasiado en pruebas de independencia, que dependen a su vez de la cantidad disponible de datos.

Los métodos basados en score exploran el espacio de CPDAG's y minimizan un score global. La exploración se realiza en términos de operaciones sobre grafos, por ejemplo añadir o remover aristas. El algoritmo Greedy Equivalent Search (GES) de **?** encuentra la estructura óptima en términos del BIC.

Los métodos híbridos combinan las técnicas anteriores. Por ejemplo el Max-Min Hill Climbing Algorithm de **?** primero construye un esqueleto utilizando pruebas de independencia condicional, y luego le da orientación a las aristas utilizando un método greedy sobre un score bayesiano.

2.8.2. Explotando la asimetría entre causa y efecto

Los métodos anteriores no toman en cuenta la asimetría entre causas y efectos. La intuición detrás (**?**) está dada por el siguiente ejemplo: Si consideramos el modelo causal funcional $Y = X + U$, entonces los métodos anteriores basados en grafos no pueden darle orientación a la arista $X - Y$, pues tanto $X \to Y$ como $Y \to X$ son Markov equivalentes.

? infiere la dirección causal al construir dos modelos Bayesianos generativos: uno para $X \to Y$ y otro para $Y \to X$. La dirección causal es determinada a partir del modelo generativo que mejor ajuste a los datos. Otros métodos, como el propuesto por **?** se basan en que si $X \to Y$, entonces la probabilidad marginal de la causa $P(X)$ es independiente del mecanismo causal $P(X|Y)$; por lo tanto, estimar $P(Y|X)$ a partir de $P(X)$ no debería poderse, pero estimar $P(X|Y)$ a partir de $P(Y)$ sí debería ser posible.

2.8.3. Como problema de Aprendizaje

Inspirados en los concursos de Kaggle de **?**, **?** plantean el problema de inferencia causal como un problema de clasificación de medidas de probabilidad. Los datos provistos por el concurso son 16,200 parejas de variables (X_i, Y_i). Cada par consta de una realización (x_i, y_i) y una etiqueta l_i que indica el valor de la verdadera relación causal; es decir, l_i indica uno de los siguientes casos: $X \to Y$, $Y \to X$, $X_i \perp Y_i$, $X \leftrightarrow Y$.

Notemos que este método tiene como punto débil que depende de la representatividad de la muestra de entrenamiento y que supone la existencia de una distribución madre

2.8.4. Modelos Causales Funcionales Neuronales

? proponen un esquema que consiste en tres pasos: primero, identificar el esqueleto según propiedades de independencia estándar; luego, orientar las aristas utilizando métodos bivariados estándar y finalmente hacer una optimización sobre todo el grafo examinando relaciones locales. Este último paso se realiza al aprender estos modelos locales como redes neuronales generativas; es decir, se modela la distribución $P(X|X_{Pa_i}, U_i)$ de un nodo condicional en sus padres (causas) con una red neuronal.

Formalmente, se modela cada ecuación

$$X_i = f_i(X_{Pa_i}, U_i)$$

como una red neuronal generativa global, en la cual cada función f_i es modelada por una red neuronal generativa con una capa oculta.

Encontrar un modelo causal de este tipo requiere llevar a cabo dos tareas: estimar la estructura (el grafo) de los padres de cada variable, y las redes neuronales f

2.9. Correlación y causalidad

Dos variables aleatorias X,Y pueden estar correlacionadas en distintos casos; por ejemplo, si X causa Y, o si tanto X como Y son causadas por otra variable desconocida V. Si e conoce la existencia de V y además podemos observarla.

3. Modelos gráficos causales

Sabemos que en una red bayesiana los arcos, aunque dirigidos, no necesariamente tienen una interpretación directa, lo único que importa es que el grafo capture adecuadamente las relaciones de dependencia entre variables aleatorias. Nos gustaría que fuera el caso que $X \to Y$ significara X *causa* Y. Resulta que existe una familia de modelos causales en los cuales se *interviene* en el mundo. La idea clave es utilizar *intervenciones ideales*, las cuales se denotan $do(Z = z)$ que significa forzar a la variable aleatoria Z a tomar el valor z, denotado de manera simple como $do(z)$. Esta operación corresponde al caso en que un agente directamente modificó el estado del

mundo y no sólo observó los valores de este de manera pasiva; por esta razón es muy importante tener en mente que es distinto $P(Y|do(z), X = x)$ que $P(Y|Z = z, X = x)$. Las preguntas causales de interés son planteadas de la forma $P(X|do(z))$ y son llamadas *quries intervencionales*. Este tipo de intervenciones surgen de manera natural en muchos problemas; por ejemplo, "si el paciente recibe este tratamiento, cuál es la probabilidad de sobrevivir la enfermedad" puede ser planteada como $P(sobreivir|do(tratamiento))$.

Formalmente, un modelo gráfico causal (**?**, **?**) tiene la misma forma que una red Bayesiana; es decir, consiste en una red acíclica dirigida que codifica una distribución entre un conjunto de variables aleatorias. La diferencia radica en la interpretación de las flechas, pues cada variable aleatoria está gobernada por un modelo causal en función de sus padres en el grafo. Una intervención consiste en forzar uno de los nodos a tomar un valor con probabilidad 1 y a recortar todas las flechas que entran en ese nodo.

Es importante que se cumpla que si dos variables están correlacionadas, entonces una es padre causal de la otra y que no falten padres causales de ninguna variable.

Para poder contestar pregunta acerca de intervenciones, se necesita introducir el concepto de *grafo mutilado*

Definición 5. *Sea $\mathcal{B}$ una red sobre un conjunto de variables aleatorias y sea $Z_1 = z_1, ..., Z_k = z_k$ una instanciación de estas variables denotada por $\mathbf{Z} = \mathbf{z}$. Se define el grafo mutilado $\mathcal{B}_{\mathbf{Z}=\mathbf{z}}$ como:*

- *Cada nodo $Z_i \in \mathbf{Z}$ no tiene padres en $\mathcal{B}_{\mathbf{Z}=\mathbf{z}}$*
- *El modelo local de probabilidad condicional para $Z_i \in \mathbf{Z}$ es tal que asignal probabilidad uno al valor z_i y cero a cualquier otro valor.*
- *No se altera ningún otro nodo $X_i \notin \mathbf{Z}$*

Entonces, utilizando este concepto, tenemos que para un modelo causal $\mathbf{C}$ una intervención $P_{\mathcal{C}}(Y|do(z)$ es realmente una *querie* probabilista estándar, pero en el grafo mutilado dado por $\mathcal{C}_{Z=z}$; es decir

$$P_{\mathcal{C}}(Y|do(z), x) = P_{\mathcal{C}_{Z=z}}(Y|x).$$

3.1. Identificabilidad de modelos

Aun a pesar de desconocer la totalidad de las variables latentes causales, existen casos en los que es posible responder a preguntas causales en modelos que involucran variables latentes utilizando sólo variables observadas; este es el caso de los modelos identificables. Existen formas de simplificar queries intervencionales y convertirlas en queries que sólo requieren de datos observacionales pues existen equivalencias entre unas y otras bajo ciertas condiciones.

? exploran las condiciones para obtener identificabilidad en modelos causales funcionales.

3.2. Aprendizaje de modelos gráficos causales

Según **?**, el problema de aprender un modelo causal a partir de datos puede dividirse en varios ejes:

- En primer lugar, definir qué se entiende por un modelo causal, pues si estamos trabajando con modelos gráficos causales entonces el problema se convierte en aprender redes bayesianas, pero con la diferencia de la interpretación que se da a las flechas; por otro lado, en modelos causales funcionales se tienen parametrizaciones mucho más complejas para las variables de respuesta, por lo que se necesitan modelos más complejos como el modelo neuronal presentado anteriormente.
- Un segundo eje es determinar si ya se conoce la estructura y sólo se necesita aprender la parametrización, o si también debe ser aprendida la parametrización.

- En tercer lugar, es el tipo de datos: observados o intervencionales. Datos observados son aquellos que provienen de alguna distribución generadora y datos intervencionales es en aquellos en los que se intervino el modelo y se observó la respuesta de este. Los datos intervencionales son difíciles de obtener en la práctica, o incluso pudieran ser inmorales o ilegales.
- Finalmente, cómo tratar los factores de confusión (confounding factors). Uno puede suponer que simplemente no existen y en estos casos el problema se vuelve relativamente sencillo.

3.2.1. Aprender sin factores de confusión

El problema de aprender modelos gráficos causales sin factores de confusión es esencialmente el mismo que aprender redes bayesianas (**?**), pero hay que hacer algunas suposiciones importantes, pues existen siempre varios DAG's que representen la misma distribución, pero que tendrían distintas interpretaciones causales. Una de las suposiciones necesarias para poder aprender modelos causales a través de datos observacionales es la llamada Propiedad Causal de Markov, que es un análogo a la Propiedad de Markov para redes Bayesianas; y su formulación es la siguiente (**?**): *cada variable es condicionalmente independiente de sus no-efectos dadas sus causas directas.* El otro supuesto necesario es que las únicas independencias condicionales sean aquellas que surgen de la d-separación en el grafo causal correspondiente (*faithfulness*).

Aun con estas suposiciones, lo más que puede aprenderse es una clase de equivalencias para el grafo causal y para llevar esto a cabo existen dos familias de métodos: basados en restricciones y basados en score y son los mismos que se utilizan para aprender el CPDAG en modelos causales funcionales.

3.2.2. Aprender de datos intervencionales

En el caso en que se tienen datos provenientes de intervenciones, lo cual es lo más común en el contexto de experimentos científicos. Se utilizan métodos basados en score para aprender de este tipo de datos.

3.2.3. Aprender con variables latentes

En el caso en el que suponemos que existen variables latentes, estamos pensando que los datos fueron generados por una distribución que es *parecida*, o parcial, a la verdadera. De nuevo, existen familias de métodos basadas en restricciones y métodos basados en scores.

3.3. Redes Bayesianas causales: aplicación en genética

? intentan aprender la estructura de las relaciones entre un conjunto de variables relacionadas a un estudio de cáncer (**?**,**?**). Específicamente, los autores analizan sólo 11 variables relacionadas a un solo gen. Identifican relaciones interesantes, como por ejemplo la variable "fumador" como padre causal de la variable "cancer", aunque no apareció relación entre la variable cancer y un gen que se sabe está relacionado con el cáncer, pero esto se debe a la manera en la que fueron seleccionados los pacientes.

4. Modelos causales funcionales y ecuaciones estructurales

En otras áreas, como las ingenierías y la física, las relaciones causales se expresan como relaciones funcionales determinísticas en donde cualquier componente aleatorio se introduce en términos de variables no observadas. Esto corresponde (**?**) a la noción Laplaciana de lo aleatorio. En este contexto, un modelo causal funcional consiste en un conjunto de ecuaciones de la forma

$$X_i = f_i(X_{Pa_i}, U_i),\ i = 1, ...n$$

donde Pa_i corresponde al conjunto de variables que directamente determinan a la variable con índice i y donde U_i representa cualquier error de medición o de aspectos no observados.

Dado un modelo causal de esta forma, podemos obtener un diagrama de la siguiente manera: para cada variable pintamos un nodo, y desde cada elemento de Pa_i dibujamos una flecha hacia el nodo que corresponde a X_i. A este diagrama se le llama *diagrama causal*. Si cumple que el diagrama es acíclico, se le conoce como semi-Markoviano. Si, además de ser acíclico, los términos de error son independientes entonces se dice que es Markoviano. El siguiente Teorema muestra la conexión entre el diagrama causal y la distribución de probabilidad asociada.

Teorema 6. *Todo modelo causal Markoviano induce una distribución $P(X_1, ..., X_n)$ que satisface que cada variable X_i es independiente de todos sus no-descendientes dado sus padres Pa_i.*

Sabemos que dos DAG's que tienen el mismo esqueleto y las mismas v-estructuras son equivalentes, y a la clase de equivalencia le llamaremos CPDAG (**?**)

4.1. Intervenciones y efectos causales en modelos funcionales

Los modelos funcionales causales proveen una manera de estimar cómo cambiaría la distribución en respuesta a cambios externos. Lo que se hace en estos modelos es representar las intervenciones como alteraciones a un conjunto de funciones. El efecto total de la intervención se calcularía modificando las ecuaciones correspondientes y con este nuevo modelo calcular las nuevas probabilidades. Denotamos una intervención como $do(X = x)$. Se dice que una variable X_i es causa directa de X_j si

$$P_{X_j|Do(X_i=x_i} \neq P_{X_j|do(X_i=x')}$$

4.2. Contrafactuales en modelos funcionales

5. Extracción de relaciones causales a partir de texto

5.1. Extracción basada en patrones gramaticales

? desarrollan un método que permite identificar y extrar relaciones causa-efecto que se encuentra expresada de manera explícita en *abstracts* de artículos médicos. Los autores deciden utilizar el dominio médico pues al ser importantes las relaciones causales es más probable que estas aparezcan explícitamente. Los autores primero identifican los *marcadores lingüísticos* y con esto construyen un conjunto de *patrones causales*. Además, cada oración es pasada por un *parse tree* el cual genera una representación de la estructura sintáctica de la oración.

Como ejemplo, consideremos la oración *"Paclitaxel was well tolerated and resulted in a significant clinical response in this patient"*. Después de la creación del árbol sintáctico y del patrón gramatical, se obtiene que la causa presente en la oración es *paclitaxel* y el efecto es *a significant clinical response in this patient*.

Analizando 200 abstracts que provienen de 4 áreas médicas (depresión, esquizofrenía, enfermedades cardiacas y SIDA), construyen los patrones causales. El proceso de extracción de información consiste en hacer un *match* entre los patrones causales y el arbol sintáctico. Al aplicar los patrones causales obtenidos a un nuevo conjunto de abstracts, obtienen un 51 por ciento de accuracy al extraer causas y un 58 por ciento de accuracy al extraer efectos.

? presentan un método semi-automático para la detección de patrones causales en texto, es semi-automático pues los patrones son descubiertos de manera automática, pero la validación no lo es. El algoritmo de detección de patrones léxico-sintácticos que se refieren a causalidad consiste en dos partes. La primera parte descubre patrones léxico sintácticos que pueden expresar relaciones causales, y la segunda valida estos patrones según restricciones semánticas. Se enfocan en encontrar patrones de la forma (frase verbo causal frase) donde el verbo es un verbo causal simple en los cuales el verbo se refiere únicamente a la causa; por ejemplo, *Temblores generan tsunamis*. Como las frases obtenidas pueden ser muy largas, pueden aportar información no deseada. Para resolver esto **?** utiliza WordNet para reducir el tamaño de las frases obtenidas.

? desarrollan un sistema para la adquisición de conocimiento causal a partir de texto. El sistema propuesto identifica oraciones que especifican de manera explícita relaciones causales y extrae de ellas los patrones causales. El sistema incorpora información gramatical tal como conectores, conjunciones, disjunciones y negaciones y codifican todo esto en redes bayesianas.

Para la detección de patrones causales, los autores se enfocan en trabajar con oraciones en las cuales ambos *eventos* aparezcan como frases (noun phrases, ¿traducción?). Un ejemplo de patrones causales identificados por el sistema es *"Anemia are caused by excessive hemolysis"*.
Para la construcción de las redes, se toma como input un texto *tokenizado*, etiquetado con *part-of-speech tagging* y *parsed* para encontrar expresiones verbales. Posteriormente, el sistema estima las probabilidades condicionales de los efectos dadas las causas mediante el conteo de frecuencias de eventos similares.

Para la evaluación, hicieron una evaluación subjetiva que consistió en comparar la estructura de la red generada por el sistema y la estructura creada manualmente al observar los patrones causales en el texto. Se analizaron 7 textos y en general se encontró un empate entre las redes con una precisión de 60 por ciento, pero sin ningún valor de Recall, por lo que no es fácil interpretar sus resultados. Según **?**, una de las razones por las cuales se ubtuvo baja precisión es que basar los nodos en las frecuencias no es una buena idea, pues la cantidad de texto disponible en general es tan grande que estas frecuencias no pueden ser representativas.

? menciona que muchos de estos métodos están más enfocados en cómo extraer las relaciones que en cómo usarlas. Es importante no sólo poder concluir que dos términos están relacionados causamente, sino cómo.

5.2. Extracción basada en patrones de co-ocurrencias

En vez de utilizar patrones gramaticales, **?** y **?** analizan la co-ocurrencia de términos estadísticos para construir redes a partir de texto. Ellos definen un "statistical term" como *"Statistical terms are expression of the measurements of statistics to watch the movements of phenomena; birth rates, public approval rating of the Cabinet and so on."*; es decir, un término que sea importante a un sistema. El enfoque de su trabajo es construir redes causales entre los statistical terms. Lo que hacen es primero extraer relaciones de co-ocurrencia entre términos estadísticos y posteriormente extraer relaciones causales entre ellos. Los autores se enfocan en texto japonés, el cual cumple que tiene ciertos patrones en los sufijos utilizados, y estos patrones son utilizados al extraer los términos estadísticos. Luegom estos términos son clasificados según su nivel de abstracción. Posteriormente, construyen una red de términos estadísticos al considerar a dos de ellos co-ocurrentes si aparecen juntos en un párrafo. Se remueven los términos que son co-ocurrentes con muchos otros términos.

Los autores evaluan de manera separada cada una de las etapas: extracción de términos, clasificación de términos, la construcción de redes y la clasificación de redes.
La mayor desventaja de esto es que la dirección de la relación causal no puede ser inferida de la co-ocurrencia.

Por otro lado, **?** extrajeron términos de artículos financieros (en japonés) y luego considerarón que estos términos afectaban el desempeño de los negocios basados en co-ocurrencias con expresiones clave como "es bueno". Su definición de causalidad depende de las expresiones clave.

5.3. Modelos Gráficos para extracción de relaciones

A partir de un artículo que trata específicamente sobre una entidad claramente indentificable (la biografía de Benito Juarez por ejemplo) **?** intentan identificar entidades y descubrir relaciones semánticas entre ellas. Resulta evidente que en el documento aparecerán otras entidades, pero se hace el supuesto de que no están relacionadas entre sí. En este trabajo, todas las clases entidades están pre-definidas así como las relaciones; esto es, se sabe que en el documento aparecerán años, lugares, países y que una entidad de tipo persona habrá nacido en un país en cierto año. Para formular el problema, se tiene $\mathbf{x} = \{x_1, ..., x_N\}$ una sucesión observada de *tokens*. Denotamos como s_p la entidad principal. Sea $\mathbf{s} = \{s_1, ..., s_L\}$ una segmentación en la cual cada s_i es una tripleta $\{\alpha_i, \beta_i, y_i\}$ con α_i es una posición inicial, β_i posición final y y_i una etiqueta asignada.

Sea r_{pn} la asignación de relaciones entre la principal s_p y un candidato a entidad secundaria s_n y sea $\mathbf{r}$ el conjunto de asignación de relaciones para la $\mathbf{x}$. Dada una observación $\mathbf{x}$, queremos encontrar $\mathbf{y}^*$ tal que

$$\mathbf{y}^* = \arg\max_y p(\mathbf{y}|\mathbf{x})$$

donde $\mathbf{y} = \{\mathbf{r}, \mathbf{s}\}$.

Se define una distribución condicional conjunta para $\mathbf{s}, \mathbf{x}, \mathbf{r}$ en un modelo grafico no dirigido $\mathcal{G}$ tal que se pueden dividir los factores de $\mathcal{G}$ en tres grupos: ϕ^S, ϕ^R, ϕ^∇, que son el potencial de segmentación, el potencial de relaciones, y el potencial conjunto de segmentación-relación.

Por el Teorema de Hammersley-Clifford, la condicional conjunta $P(\mathbf{y}|\mathbf{x})$ se factoriza como

$$P(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \prod_{C_S} \phi^S(i, \mathbf{s}, \mathbf{x}) \prod_{C_R} \phi^R(r_{pm}, r_{pn}, \mathbf{r}) \prod_{C_\nabla} \phi^\nabla(s_p, s_j, \mathbf{r})$$

Se hace el supuesto de que las funciones potencial se factorizan según un conjunto de atributos y un conjunto de pesos correspondientes.

Incorporando estos features, tenemos que

$$P(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \exp\left(\sum_{i=1}^{|s|} \sum_{k=1}^{K} \lambda_k g_k(i, s, x) + \sum_{m,n}^{M} \sum_{w=1}^{W} \mu_w q_w(r_{pm}, r_{pn}, r) + \sum_{j=1}^{L} \sum_{t=1}^{T} \nu_t h_t(s_p, s_j, r) \right)$$

Dada una muestra $\mathcal{D} = (\mathbf{x}_i, \mathbf{y}_i)_{i=1}^N$ queremos estimar los parámetros $(\lambda_k, \mu_w, \nu_t)$.

La log-verosimilitud regularizada es

$$\mathcal{L} = \log[\Phi(r, s, x)] - \log[Z(x)] - \sum_{k=1}^{K} \frac{\lambda_k^2}{2\sigma_\lambda^2} - \sum_{w=1}^{W} \frac{\mu_w^2}{2\sigma_w^2} - \sum_{t=1}^{T} \frac{\nu_t^2}{2\sigma_\nu^2}$$

Donde,

$$\Phi(r, s, x) = \exp(\sum_{i=1}^{|s|} \sum_{k=1}^{K} \lambda_k g_k(i, s, x) + \sum_{m,n}^{M} \sum_{w=1}^{W} \mu_w q_w(r_{pm}, r_{pn}, r) + \sum_{j=1}^{L} \sum_{t=1}^{T} \nu_t h_t(s_p, s_j, r))$$

Tenemos que $\mathcal{L}$ es concava por lo que es fácil maximizarla.

Para calcular la Asignacion Más Probable, hayue encontrar

$$\mathbf{y}^* = \arg\max_y p(\mathbf{y}|\mathbf{x})$$

La cual no se puede resolver de manera exacta, se necesitan métodos aproximados. El algoritmo que se propone en el artículo se llama Collective Iterative Classification y la idea detrás de este es decodificar las variables ocultas objetivo basados en asignar etiquetas a las variables muestreadas. Esto se hace mediante un proceso iterativo de dos pasos; primero, en una etapa de bootstrapping se predice una etiqueta inicial para una x_i dado el modelo ya entrenado. Luego, en la segunda etapa, conocida como clasificación iterativa, se re-estima la asignación a x_i varias veces, tomándolas a partir de la asignación inicial.

Los autores prueban su algoritmo con datos provenientes de Wikipedia: sus datos consisten en 441 páginas de Wikipedia. De estos datos, fueron etiquetados a mano 7740 entidades en 8 categorías. Además, se extrajeron 4700 relaciones de las cuales las más frecuentes fueron trabajo, visió, nació, miembro de, nació en día.

El modelo superó los existentes tomando como medida de calidad la medida F.

6. Métodos Estadísticos Para Extracción de Relaciones Causales

6.1. Statistical Relational Learning

En el contexto de inferencia causal, se estudian las interacciones entre variables que puedan causar efectos en la otra. Para esto, es necesario identificar los objetos y las relaciones entre ellos. Para esto, el área de Statistical Relational Learning ofrece una serie de métodos que modelan dependencias entre entidades relacionales. En esta sección veremos distintas técnicas existentes en esta área y sus aplicaciones a inferencia causal.

Un dominio de interés consiste en objetos (entidades), sus atributos, y las relaciones entre sí de los objetos. El modelo DAPER de **?** consiste en clases de entidades, clases de relaciones, clases de atributos y clases arco.

6.1.1. Aplicación de Campos Aleatorios Condicionales a Extraccion de informacion

? analizaron 753,459 abstracts de journals médicos, de los cuales extrajeron 6580 interacciones entre 3,737 proteínas. Para ello, utilizaron un algoritmo de tres etapas: la primera etapa, consiste en identificar nombres de proteínas mediante un clasificador basado en Conditional Random Fields; la segunda, identificar interacciones a través de co-ocurrencias; posteriormente, filtrar estas interacciones con un clasificador Bayesiano para obtener interacciones válidas. Con esto, se obtiene una red que consiste en 31609 interacciones entre 7748 proteínas

Para la etapa de identificación de nombres, se utilizó un Conditional Random field, cuyo desempeño fue evaluado en un conjunto de 200 abstracts médicos etiquetados y en 750 abstracts etiquetados a mano. Para extraer los nombres de proteínas, se utilizó este clasificador en el conjunto de 753,459 abstracts de Medline que contuvieran la palabra "human".

Para obtener las interacciones, se midió la co-citación entre los nombres de proteínas y luego se enriquecieron estos pares con interacciones físicas (reales) entre proteínas utilizando un filtro Bayesiano. Primero, se contó el número de veces que en un abstract aparece un par de proteínas, y con eso se calculó la probabilidad de co-citación. Con esto se obtienen 15,000 interacciones, pero se pueden dar co-citaciones por otras razones además de las relaciones reales, se aplicó un filtro bayesiano que mide la verosimilitud de los abstracts que citan un par de proteínas para discutir sus interacciones reales. El clasificador asigna un *score* a cada uno de los abstracts que citan esas proteínas. Utilizando un umbral de score, se obtienen 6,580 interacciones entre 3,737 proteínas.

Al combinar estas interacciones con otras 26,280 interacciones provenientes de otras fuentes, se obtiene el total mencionado.

Por otro lado, **?** plantean el problema de extracción de relaciones como un problema de etiquetado de secuencias, algo similar a lo realizado por **?**. **?** definen un modelo de la forma

$$P_\Lambda(y|x) = \frac{1}{Z} \prod_{c \in C} \varphi_c(y_c, x_c, \Lambda),$$

donde las φ son potenciales parametrizados por Λ.

Los autores tomaron 1,127 párrafos de 271 artículos biográficos de Wikipedia y etiquetaron 4,701 instancias de relaciones. Para identificar a las entidades importantes dentro de un artículo, utilizaron el hecho de que estas suelen tener un *hyperlink*

6.1.2. Aplicación de RMN's a Extraccion de informacion

? utilizan Relational Markov Networks para mejorar el desempeño de los CRF's en el problema de identificar nombres de proteínas en abstracts de journals médicos.

La técnica que se propone es que dada una colección de documentos D, asociemos a dada documento $d \in D$ un conjunto de entidades $d.E$. Cada entidad$e \in d.E$ queda caracterizada por un conjunto de atributos booleanos $e.F$. A cada documento se le asocia un grafo de factores, el cual es un grafo bipartita que contiene dos tipos de nodos: nodos variables, que corresponden a las etiquetas de todas las entidades candidato en el documento y nodos potenciales, que modelan las correlaciones entre dos o más atributos. Para cada correlación posible, un nodo potencial es crreado tal que está ligado a todos los nodos variables involucrados.

Prueban su algoritmo en dos conjuntos de datos previamente etiquetados a mano. El primer conjunto consiste en 200 abstracts y el segundo en 225. Ambos conjuntos de datos fueron previamente etiquetados a mano, además de etiquetados para POS. En ambos conjuntos de datos, existe una mejora ligera respecto a los CRF's en términos de la medida F.

6.1.3. Aprendizaje de Modelos Causales a partir de Datos Relacionales

? definen la *abstract ground graph* (AGG) y se desarrolla la noción de d-separación relacional. A partir de estas nociones teóricas, **?** extienden el algoritmo PC, el cual identifica todas las posibles orientaciones de dependencias causales, a datos relacionales. Utilizando la d-separación relacional, introducen restricciones que orienta dependencias bivariadas hasta en un 72 % más. Se prueba teóricamente que con esta nueva regla, además de extensiones a datos relacionales del algoritmo PC, se obtiene un método *sound and complete* para la extracción de relaciones causales. Este algoritmo nuevo se llama RCD (relational causal discovery).

? notan que la demostración de **?** requiere que la AGG sea un DAG quie represente exactamente todas las flechas que aaparecerían en todas las posibles ground graphs, y en un trabajo anterior (**?**) habían notado que existen casos en los cuales la d-separación de la AGG no captura las independencias condicionales que se cumplen en el modelo relacional causal. En su artículo, proponen un algoritmo basado en condiciones más débiles llamado RCD-light.

6.2. Inferencia Bayesiana No Paramétrica

La rama no-paramétrica de la inferencia bayesiana puede considerarse como una extensión a espacios infinito-dimensionales de la inferencia bayesiana clásica. Esto permite tener modelos que permiten capturar una complejidad mucho mayor en los datos. El objeto principal de la inferencia bayesiana no paramétrica es el Proceso Dirichlet, que es una extensión a espacios infinito-dimensionales de la distribución Dirichlet clásica.

6.2.1. Proceso Dirichlet

Definición 7. *Sea H una medida finita definida sobre un espacio $\mathcal{X}$ y $\alpha \in \mathbb{R}$. Una medida aleatoria G sobre $\mathcal{X}$ se dice que es un Proceso Dirichlet de parámetros (α, H) si para cada partición finita medible $\{A_1, ..., A_k\}$ de $\mathcal{X}$ se tiene que la distribución conjunta de $(G(A_1), ..., G(A_k))$ es Dirichlet de parámetros $(\alpha H(A_1), ..., \alpha H(B_k))$. A la medida H se le conoce como medida base y a α se le conoce como parámetro de concentración. Se denota como*

$$G \sim \ DP(\alpha, H).$$

Los parámetors H y α juegan roles intuitivos en la definición. La medida base se puede entender, básicamente, como la *media* del Proceso; es decir, para cada conjunto medible $A \subseteq \mathcal{X}$ se tiene que $\mathbb{E}[G(A)] = H(A)$. Por otro lado, el parámetro de concentración es un tipo de *varianza inversa*, pues $V[G(A)] = H(A)(1-H(A))/(\alpha+1)$. Es decir, mientras más grande sea α, entonces menor es la varianza y el DP tenderá a concentrarse más en torno a H.

6.2.2. Distribución Predictiva y urnas de Blackwell-McQueen

Consideremos $G \sim \text{DP}(\alpha, H)$ y $\theta_1, ... \sim G$. Como $\theta_{n+1}|G, \theta_1, ..., \theta_n \sim G$, entonces para $A \subseteq \mathcal{X}$ tenemos que

$$\begin{aligned} P(\theta_{n+1} \in A|\theta_1, ..., \theta_n) &= \mathbb{E}[G(A)|\theta_1, ..., \theta_n] \\ &= \frac{1}{\alpha + n}\left(\alpha H(A) + \sum_{i=1}^{n} \delta_{\theta_i}(A)\right) \end{aligned}$$

Por lo tanto, marginalizando G, vemos que la distribución posterior base dados $\theta_1, ..., \theta_n$ es la distribución predictiva de θ_{n+1}.

$$\theta_{n+1}|\theta_1, ..., \theta_n \sim \frac{1}{\alpha + n}\left(\alpha H + \sum^{n} \delta_{\theta_i}\right)$$

La secuencia de distribuciones predictivas para $\theta_1, \theta_2, ...$ se conoce como el esquema de urnas de Blackwell-McQueen; la interpretación es la siguiente: cada valor en $\mathcal{X}$ es un color, y cada realización $\theta \sim G$ son pelotas y el valor obtenido es el color de la pelota. Además, tenemos una urna en la cual se han colocado las bolas observadas. Al principio, no existen bolas en la urna y muestreamos un color de H; es decir, $\theta_1 \sim H$, pintamos la bola con ese color y la echamos en la urna. Luego, en el paso $n+1$, tomaremos un nuevo color con probabilidad $\alpha/(\alpha+n)$ o tomaremos un nuevo color, la pintaremos así y la echaremos en la urna con probabilidad $n/(n+\alpha)$.

6.2.3. Proceso Dirichlet como Prior en problemas de Clustering

La propiedad de que $G \sim DP$ sea discreto ofrece aplicaciones de clusterización. Por el momento, supongamos que H es suave. Entonces, como los valores de las realizaciones son repetidos, sean $\theta_1^*, ..., \theta_m^*$ los valores únicos de las realizaciones $\theta_1, ..., \theta_n$ y sea n_k el número de repeticiones de θ_k^*. Entonces, la distribución predictiva se puede escribir como

$$\theta_{n+1}|\theta_1, ..., \theta_n \sim \frac{1}{\alpha + n}\left(\alpha H + \sum_{k=1}^{m} n_k \delta_{\theta_k^*}\right).$$

Esto es, el valor de θ_k^* aparecerá en θ_{n+1} con una probabilidad proporcional al número de veces que ha aparecido. Entonces, mientras más grade sea n_k, mayor será la probabilidad de que se vuelva aun mayor. Notemos que los valores únicos de las realizaciones $\theta_1, ..., \theta_n$ inducen una partición aleatoria del conjunto $[n] = \{1, ..., n\}$ si pensamos el cluster k como aquel que contiene las θ_i que toman el valor θ_k^*, tenemos que clusters grandes crecen más rápido.

Podemos invertir este proceso y empezar con una distribución sobre particiones.

6.2.4. Dirichlet Enhanced Relational Learning

? aplican un modelo jerárquico Bayesiano no-paramétrico para la extracción de relaciones. La ventaja de usar un modelo jerárquico es que los parámetros pueden estar *personalizados* para entidades. En este trabajo se aplica el modelo al contexto médico, en el cual se tienen como entidades hospitales, pacientes, diagnósticos y procedimientos. Se sabe que la existencia de un diagnóstico o un procedimiento es dependiente de las características de un paciente; por otro lado, las características de los hospitales se modelan como inciertas.

Un modelo Bayesiano consiste en un parámetro θ y una asignación inicial de probabilidad sobre este, la cual refleja nuestra incertidumbre inicial. Supondremos que esta distribución inicial cuenta con hiperparámetros h, de modo que sea $P(\theta|h)$. La verosimilitud de los datos $P(D|\theta)$ refleja cómo son generados estos dado el parámetro. EN el caso en el cual se asume que los datos provienen de distintos grupos es natural pensar en asignar un parámetro por grupo, lo cual lleva a la definición de modelos jerárquicos, en los cuales el modelo tiene la siguiente fforma

$$P(h)\prod P(\theta_i|h)P(D_i|theta_i).$$

Si se aplicara el modelo DAPER al contexto médico, se tiene que los parámetros e hiperparámetros que especifican distribuciones condicionales como atributos globales. Esto implica que las probabilidades de un procedimiento son idénticas para todos los pacientes con la misma dolencia inicial. Luego, los procedimientos son modelados de manera independiente de manera que procedimientos anteriores en una persona no afectlan la selección de uno nuevo. Estas son suposiciones no realistas.

Para mejorar esta situación, se asume que la probabilidad de un procedimiento es una distribución multinomial con un prior Dirichlet. En particular, para cada paciente, se asume que un vector individual de parámetros $\theta_{s|pc,pa}$ especifica la probabilidad de un procedimiento para un paciente *pa* con dolencia inicial *pc*. Este vector es generado por una distribución Dirichlet de parámetros $h_{pc} = \{\tau_{pc}, \alpha_{pc}\}$, la cual puede escribirse como

$$\text{Dir}(\theta_{\cdot|pc,pa}|\tau_{pc}, \alpha_{pc}) = \frac{1}{C}\prod_{k=1}^{K} \theta_{k,s|pc,pa}^{\tau_{pc}\alpha_{k,pc}-1}.$$

En un modelo jerárquico bayesiano cada paciente obtiene sus propias probabilidades de procedimiento y comparte con los demás un parámetro a priori.

En más de un caso, se necesita asumir que la prior tiene una forma muy flexible, por lo que es necesario recurrir a priors no paramétricas.

Podemos tomar la distribución a priori como una realización del Proceso Dirichlet; es decir,

$$G_{pc} \sim \text{ DP } (G_0, \alpha_0)$$

donde G_0 es la medida base y α_0 el parámetro de concentración, el cual especifica nuestro nivel de certidumbre sobre la prior.

Luego, el parámetro $\theta_{\cdot|pc,pa}$ es una realización de G_{pc}.

Otra manera de escribir el Proceso Dirichlet es mediante su representación *stick breaking* (**?**) según la cual

$$G_{pc} = \sum \pi_{l,pc}\delta_{\theta^*_{l,pc}} \text{ ; } \theta^*_{l,pc} \sim G_0$$

$$\pi'_{l,pc} \sim \text{ Beta } (1, \alpha_0) \text{ ; } \pi_{l,pc} = \pi'_{l,pc}\prod_{k=1}^{l-1}(1 - \pi'_{k,pc}).$$

Notemos que aunque la distribución base pueda ser continua, el DP es siempre discreto (con probabilidad 1).

Tradicionalmente, el aprendizaje en el contexto Bayesiano no paramétrico se lleva a cabo via muestreo de Gibbs, cuyas variaciones más comunes son la urna de Polya o el Proceso del Restaurante Chino (**?**). En el artículo, se lleva a cabo un proceso más eficiente (**?**.)

El objetivo es estimar G_{pc} para cada posible *pc* en la base de datos, utilizando la verosimilitud marginal:

$$\hat{G}_{pc} = \text{ argmax DP } (G|G_0, \alpha_0) \prod_{\{pa\}_{pc}} \int \text{ Multinomial } (\{pr\}_{pa}|\theta_{\cdot|pc,pa})d\theta_{\cdot|pc,pa}$$

donde $\{pa\}_{pc}$ es el conjunto de pacientes tales que tienen la misma dolencia inicial y $\{pr\}_{pa}$ es el conjunto de procedimientos para el paciente *pa*. Resulta imposible calcular la verosimilitud, por lo que se realiza una aproximación de campo medio:

$$G_{pc} \approx \sum^{|pa_{pc}|} \pi_{pa,pc}\delta_{\theta^*_{\cdot|pa,pc}}.$$

El proceso de aprendizaje se divide en dos etapas:

- Calcular la ubicación de $\theta^*_{\cdot|pa,pc}$
- Estimar los pesos $\pi_{pa,pc}$

Para el primer paso, se utiliza la aproximación MAP de $\theta_{\cdot|pa,pc}$ que es:

$$\theta^{MAP}_{\cdot|pa,pc} = \text{ arg max } P(\theta_{\cdot|pa,pc}|\{pr\}_{pa})$$

Y se utiliza como prior la Dirichlet mencionada antes.

Para el segundo paso, es necesario hacer la siguiente suposición:

$$\hat{P}(\theta_{\cdot|pa,pc}|\{pr\}_{pa}) \approx q_{pa}(\theta_{\cdot|pc,pa}) = \sum_{\{\tilde{pa}\}_{pc}} \xi_{\tilde{pa},pc,pa}\delta_{\theta^{MAP}_{\cdot|pc,pa}}$$

donde $\xi_{\tilde{pa},pc,pa} \geq 0$ son parámetros variacionales tales que $\sum_{pa} \xi_{\tilde{pa},pc,pa} = 1$.

Como paso E se obtiene que:

$$\xi^t_{\tilde{pa},pc,pa} = \frac{P(\{pr\}_{\tilde{pa}}|\delta_{\theta^{MAP}_{\cdot|pc,pa}})\hat{G}^{(t)}_{pc}(\delta_{\theta^{MAP}_{\cdot|pc,pa}}))}{\sum_{\tilde{pa}} P(\{pr\}_{\tilde{pa}}|\delta_{\theta^{MAP}_{\cdot|pc,pa}})\hat{G}^{(t)}_{pc}(\theta^{MAP}_{\cdot|pc,pa}))}.$$

Y el paso M queda como

$$\hat{G}^{(t+1)}_{pc}(\theta_{\cdot,pc,pa}) = \frac{\alpha_0 G_0(\theta_{\cdot,pc,pa}) + \sum_{\{pa\}_{pc}} \xi_{pc,pa}\delta_{\theta^{MAP}_{\cdot|pc,pa}}}{\alpha_0 + |\{pa\}_{pc}|}$$

6.2.5. Learning Systems of Concepts with an Infinite Relational Model

Suponiendo que tenemos una o más relaciones que involucran uno o más tipos, **?** definen el Infinite Relational Model, el cual busca partir cada tipo (type) en clusters, de manera que una buena partición permita que relaciones entre entidades se puedan predecir a partir de la asignación de clusters. Por ejemplo, si tenemos m relaciones que involucran n tipos, denotamos R^i la i-ésima relación, T^j la j-ésima relación y z^j un vector de asignación de clusters. Lo que se busca es inferir la asignación de clusters, y ultimadamente la posterior $P(z^1, ..., z^n|R^1, ..., R^m)$. Para especificar esta distribución, se define un modelo generativo para las relaciones y la asignación de clusters

$$P(R^1, ..., R^m, z^1, ..., z^n) = \prod_{i=1}^{m} P(R^i|z^1, ..., z^n) \prod_{j=1}^{n} P(z^j).$$

Estamos asumiendo que las relaciones son condicionalmente independientes dada la asignación de clusterss, y que la asignación de clusters es independiente entre tipos. Para que el IRM sea capaz de descubrir el número de clusters en el tipo T, se utiliza un prior que asigna probabilidad a todas las posibles particiones de T. Un prior razonable debería descubrir sólo tantos clusters como sugeridos por los datos. A partir de trabajos previos en Bayesiana No paramétrica (**?**) utiliza una distribución sobre particiones inducida por un CRP (Chinese Restaurant Process, **?**).

La intuición detrás de este proceso, según describe **?** es que en un restaurante chino hay un número infinito de mesas, cada una de las cuales puede sentar un número infinito de clientes. El primer cliente entra y se sienta en la primera mesa, el segundo cliente entra y decide si sentarse en la primera mesa o en otra; en general, el n-ésimo cliente se sienta en una mesa ocupada con una probabilidad proporcional al número de clientes sentados ahí.

Entonces, la manera de constuir clusters utilizando el CRP es partir de un cluster con un solo objeto, e ir

añadiendo objetos de modo que la probabilidad de caer en cierto cluster sea proporcional a los elementos en él. De esta manera, la distribución de clusters para el objeto i, dada la asignaclión de los objetos $1, ..., i-1$ es:

$$P(z_i = a|z_1, ..., z_{i-1}) = \begin{cases} \frac{n_a}{i-1+\gamma} & \text{if } n_a > 0 \\ \frac{\gamma}{i-1+\gamma} & a \text{ es un nuevo cluster} \end{cases}$$

Donde n_a es el número de objetos asignado al cluster a y γ es un parámetro. La distribución en z inducida por el CRP es intercambiable; es decir, no importa el orden en que los objetos llegan a los clusters. Como nuevos objetos siempre pueden ser asignados, el IRM efectivamente tiene acceso a un número infinito de clusters, de ahí el nombre del Infinite Relational Model.

Para generar Relaciones a partir de las asignaciones de clusters se asume que las relaciones son binarias, pero se puede extender a datos continuos y frecuencias. Como ejemplo, veremos primero el caso con un solo tipo T y una relación $R : T \times T \to \{0, 1\}$. T puede ser, por ejemplo, un conjunto de personas y R una relación de amistad. El modelo generativo para este problema es

$$\begin{aligned} z|\gamma &\sim CRP(\gamma) \\ \eta(a, b)|\beta &\sim Beta(\beta, \beta) \\ R(i, j)|z, \eta &\sim Bernoulli(\eta(z_i, z_j)) \end{aligned}$$

Esto es, estamos suponiendo que la relación R está generada a partir de dos estructuras latentes: una partición z y una matriz de parámetros η. La entrada $R(i, j)$ se genera al lanzar una moneda con sesgo $\eta(z_i, z_j)$ donde z_i, z_j son las asignaciones de cluster de las entidades i,j. El parámetro $\eta(a, b)$ especifica la probabilidad de que exista un link entre las entidades $i \in a$, $j \in b$. El IRM invierte ese modelo para descubrir la z y la η que mejor explican R.

Estamos asumiendo que la tendencia de una entidad a participar en relaciones está dada por su asignación de cluster.

Para hacer inferencia, notamos que en el ejemplo anterior se utilizó una prior conjugada, es sencillo calcular $P(R|z)$

$$P(R|z) = \prod_{a,b} \frac{Beta(m(a, b) + \beta, \bar{m}(a, b) + \beta}{Beta(\beta, \beta)}$$

donde $m(a, b)$ es el número de pares (i, j) tales que $i \in a$, $j \in b$ y $R(i, j) = 1$ mientras que $\bar{m}(a, b)$ son las parejas donde $R(i, j) = 0$.

Para muestrear de la posterior $P(z|R) \propto P(R|z)P(z)$ se pueden utilizar técnicas MCMC o al buscar la moda de la distribución.

Para la evaluación del modelo, los autores generan datos sintéticos. Consideran conjuntos de datos de tres formas distintas; el primer sistema, $S1$, tiene sólo dos tipos $T1, T2$ y una sola relación $R : T1 \times T2 \to \{0, 1\}$. El sistema $S2$ consiste en 4 tipos y tres relaciones binarias mientras que $S3$ consiste de 3 tipos y una sola relación ternaria. Resulta que en este contexto, el modelo captura el número de clusters de manera precisa.

Este método es excelente para clusterizar de manera no paramétrica un conjunto de datos relacionales, pero requiere identificar previamente las entidades y las relaciones.

6.2.6. Learning Infinite Hidden Relational Models

? extienden la expresividad de los modelos relacionales al introducir para cada entidad una variable latente con un número infinito de estados que proviene de un Dirichlet Process Mixture. Debido a que el número de features de las cuales un atributo puede depender, resulta más conveniente introducir para cada entidad una

variable latente que es padre de los atributos de la entidad y de los atributos de las relaciones en las que participa la entidad; a esto se le conoce como hidden relationship model.

Como cada entidad puede tener un distinto número de estados en sus variables latentes, es deseable que este número de estados latentes se determine de manera automática según los datos; esto es posible utilizando una mezcla de Proceso Dirichlet, que son modelos de mezcla con un número infinito de componentes de mezcla, pero que según los datos la complejidad se ajusta de manera automática. A la combinación de un hidden relational model con un DPm se le conoce como Infinite Hidden Relationship Model que generaliza el modelado jerárquico bayesiano noparamétrico al caso de modelos relacionales.

Como aplicación, los autores evalúan su modelo en un conjunto de datos sobre películas y buscan predecir las preferencias de los usuarios. Para este conjuto de datos, las clases de entidades son Usuario y Película

6.3. Modelos Jerárquicos Bayesianos

Los Modelos Jerárquicos Bayesianos (HBM) son modelos probabilistas que están definidos por *niveles*, en los cuales cada nivel representa un mayor grado de abstracción. Un ejemplo sencillo de esto puede ser un modelo de regresión de ingreso a nivel nacional en el cual existan parámetros por estado y por municipio.

En el contexto causal, lo que está detrás de un HBM es que al aprender causalidad se incorpora conocimiento teórico previo y datos observables (**?**, **?**). Esto es, un HBM es una distribución de probabilidad entre supuestos teóricos, modelos causales y datos (**?**).

? proponen un framework que incorpora inferencia basada en covarianzas así como conocimiento adquirido. Se enfocan en aprender *forma funcional* de las relaciones causales. Las contribución principal del artículo es: se muestra que el problema de aprender la forma funcional de una relación causal puede ser formalizada utilizando modelos jerárquicos bayesianos; además, se muestran una serie de experimentos que ponen a prueba las predicciones cualitativas hechas por este modelo.

? presentan un overview general sobre los modelos jerárquicos bayesianos en el contexto causal. En el artículo exponen las ventajas de utilizar estos modelos para realizar inferencia causal, así como sus limitaciones. Una limitación mencionada es que dependen considerablemente de conocimiento previo. Otra, que existe evidencia que la representación que las personas tienen de modeos causales no corresponde a modelos gráficos causales (redes bayesianas) pues existen experimentos en los cuales la propiedad Markoviana no se cumple. Finalmente, argumentan que los HBM's y las redes bayesianas no modelan los procesos cognitivos subyacentes al apredizaje causal, aunque existen evidencias de que las personas sí llevan a cabo un tipo de actualización Bayesiana

7. Aprendizaje por Refuerzo

El Aprendizaje por Refuerzo (RL) (**?**) consiste en un *framework* general en el cual un *agente* aprende a llevar a cabo una tarea a través de interactuar con su entorno. Estas interacciones provocan cambios en el entorno, los cuales a su vez afectarán el comportamiento del agente. El comportamiento del agente está caracterizado en términos de acciones y las acciones tomadas por este afectan el *estado* del ambiente. A su vez, el ambiente devuelve información al agente en términos de una recompensa o ganancia según el estado en el que se encuentre. El objetivo del agente es encontrar una serie de acciones que lo lleven a lograr su objetivo final. La función de recompensa codifica qué es lo que se desea que el agente aprenda, pues esta función devuelve un valor para cada estado y con esto el agente asignará un valor a cada estado según la ganancia a largo plazo que un este representa.

La dinámica consiste en lo siguiente: en un tiempo t, el agente se encuentra en el estado s_t, toma una acción $a_t \in \mathcal{A}(s_t)$ de acuerdo a una *policy*, o política, y llega a un nuevo estado s_{t+1} según una probabilidad de transición $\mathcal{T}$ y recibe una recompensa r_{t+1}. El objetivo del agente es encontrar la política que maximice la recompensa.

La regla general que se sigue para la definición del ambiente (**?**) es considerar que todo aquello que no pueda ser explícitamente alterado por el agente pertenece al ambiente. La frontera agente-ambiente tiene que ver con los límites de lo que el agente puede controlar, no sobre lo que puede *conocer*. Veremos ahora con detalle las partes que conforman esta dinámica.

7.0.1. Objetivos y Recompensas

Debido a que el objetivo del agente es maximizar la recompensa en el largo plazo, la manera en la que logramos que aprenda una tarea en particular es diseñando una función recompensa adecuada. Por ejemplo, un agente que queremos que aprenda a jugar ajedrez debe ser premiado por ganar los juegos, no por lograr sub-objetivos como comer más piezas del oponente o tomar control del centro del tablero (**?**), pues podría suceder que el agente encuentre cómo lograr esto sin ganar el juego necesariamente. Es mediante la recompensa que le comunicamos al agente qué queremos de él, no cómo queremos que lo haga.

7.0.2. Definición de Cadenas de Markov

Una Cadena de Markov discreta es un proceso estocástico en el cual el estado que toma el proceso sólo depende del estado en el tiempo anterior. Esto es,

$$P(X_{t+1} \in A | X_0 = x_0, X_1 = x_1, ..., X_t = x_t) = P(X_{t+1} \in A | X_t = x_t).$$

Estos procesos quedan caracterizados por sus probabilidades de transición $p_{ij}^t = P(X_{t+1} = j | X_t = i)$. Cuando estas probabilidades son iguales para toda t, entonces se dice que la cadena es homogénea, y por esta razón podemos almacenar las probabilidades de transición en una matriz P tal que $P_{ij} = p_{ij}$.

7.0.3. Procesos de Decisión Markovianos con recompensa

Consideremos una Cadena de Markov con una ligera modificación: al tiempo t, se toma una *acción* a_t que afecta el estado del proceso al tiempo $t + 1$. En respuesta a la acción tomada, se recibe una recompensa $r_t = f(a_t)$ para alguna función f desconocida, la cual puede ser aleatoria o determinista. Los procesos de decision Markovianos con recompensa (**?**) están definidos por un espacio de estados $\mathcal{S}$, un espacio de acciones posibles $\mathcal{A}$, una función de transición f y la función recompensa r. La función transición puede ser estocástica o determinística.

7.0.4. Función valor

Mientras la función recompensa provee la recompensa inmediata de estar en un estado, la función valor asigna una valoración *a largo plazo* de un estado. Como las recompensas obtenidas a largo plazo dependen de las acciones que se sigan tomando, tenemos que definir la función valor en términos de una política. Una política es un mapeo $\pi : \mathcal{S} \times \mathcal{A} \to \pi(s, a)$. Es decir, la probabilidad de tomar la acción a estando en el estado s. La mayoría de los algoritmos de aprendizaje por refuerzo intentan estimar de una u otra forma la función valor, por lo que la definición de esta es muy importante. Formalmente, para un proceso de decisión Markoviano, definimos la *función valor* (value function en la literatura) para un estado s, y relativa a una política π, como el retorno esperado si se comienza en s y a partir de ahí se sigue la política π

$$V^\pi(s) = \mathbb{E}_\pi \left[R_t | s_t = s \right].$$

De manera similar, definimos la *función acción-valor* $Q^\pi(s, a)$ es una valoración de tomar la acción a estandoen el estado s, y después seguir la *política* π

$$Q^\pi(s, a) = \mathbb{E}[R_t | s_t = s, a_t = a].$$

Por la ley de los grandes números, podemos estimar ambas funciones V^π y Q^π si se promedian los retornos recibidos al llegar a cada estado con el número de veces que se ha llegado al estado. Esta es la idea detrás de

los llamados Métodos Monte-Carlo que serán expuestos en una sección posterior.
Notemos una propiedad interesante de la función V^π:

$$\begin{aligned}
V^\pi(s) &= \mathbb{E}_\pi\left[R_t|s_t=s\right] \\
&= \mathbb{E}_\pi\left[\sum_{k=0}^{\infty}\gamma^k r_{t+k+1}|s_t=s\right] \\
&= \mathbb{E}_\pi\left[r_{t+1}+\gamma\sum_{k=0}^{\infty}\gamma^k r_{t+k+2}|s_t=s\right] \\
&= \sum_a \pi(s,a)\sum_{s'}P(s_{t+1}=s'|s_t=s,a_t=a)\left[\mathbb{E}[r_{t+1}|s_t=s,a_t=a,s_{t+1}=s']+\gamma\mathbb{E}\left[\sum\gamma^k r_{t+k+2}|s_{t+1}=s'\right]\right] \\
&= \sum_a \pi(s,a)\sum_{s'}\mathcal{P}^a_{ss'}\left[\mathcal{R}^a_{ss'}+\gamma V^\pi(s')\right]
\end{aligned}$$

A esta última ecuación se le conoce como la Ecuación de Bellman para V^π.

También para la función $Q^\pi(s,a)$ se tiene una ecuación de Bellman:

$$Q^\pi(s,a)=\mathbb{E}_{s',a'}[r+\gamma Q^\pi(s',a')|s',a'].$$

Además, la función V^π puede ser obtenida a partir de la función $Q^\pi(s,a)$ de la siguiente manera:

$$V^\pi(s)=Q^\pi(s,\pi(s))$$

7.0.5. Soluciones y Optimalidad

Resolver un problema de aprendizaje por refuerzo consiste en (**?**) encontrar una política que, a la larga, reciba una alta recompensa. Resulta que es posible definir un orden parcial entre políticas en términos de la función valor, pues $\pi\geq\pi'$ si $V^\pi(s)\geq V^{\pi'}(s)$ para todo estado s. Para procesos de decisión Markovianos finitos existe siempre una política mejor que las demas, a la cual llamaremos *política óptima*. De hecho, pueden existir varias, pero todas ellas comparten la misma función valor, denotada como V^*. Si para una política π^* se tiene que es mejor que todas las demás, entonces podemos describir su función valor como

$$V^{\pi^*}=\max_\pi V^\pi(s).$$

Por supuesto, $V^*=V^{\pi^*}$
Análogamente, las políticas óptimas tienen la misma función Q óptima, por lo que definimos *función accion-valor óptima* como el retorno esperado de tomar la acción a en el estado s y a partir de ahí seguir una política óptima:

$$Q^*(s,a)=\max_\pi Q^\pi(s,a)=Q^{\pi^*}(s,a).$$

Los métodos modernos de aprendizaje por refuerzo intentan aprender directamente la función valor o la función Q, y una vez Q y una vez que tenemos Q^{π^*} podemos definir una política óptima de la siguiente manera:

$$\pi^*=\text{argmax}_a Q^*(s,a).$$

Además, utilizando la ecuación de Bellman vimos que se podía escribir Q en términos de V, y por lo tanto podemos escribir Q^* en términos de V^* como

$$Q^*(s,a)=\mathbb{E}[r_{t+1}+\gamma V^*(s_{t+1})|s_t=s,a_t=a].$$

La ecuación de Bellman puede escribirse sin hacer referencia a ninguna política de la siguiente manera:

$$V^*(s)=\max_a\mathbb{E}[r_{t+1}+\gamma V^*(s_{t+1})|s_t=s,a_t=a].$$

La función óptima Q^* lo que hace es maximizar sobre todas las decisiones:

$$\begin{aligned} Q^* &= r_{t+1} + \gamma \max_{a_{t+1}} r_{t+2} + \gamma^2 \max_{a_{t+2}} r_{t+3}... \\ &= r_{t+1} + \gamma \max_{a_{t+1}} Q^*(s_{t+1}, a_{t+1}). \end{aligned}$$

Formalmente:

$$Q^*(s,a) = \mathbb{E}_{s'}[r + \gamma \max_{a'} Q^*(s', a')|s, a]$$

7.1. Métodos básicos de solución: Model-based

7.1.1. Programación Dinámica

Los métodos de programación dinámica tienen muy buen fundamento teórico, pero requieren una especificación completa del ambiente.

- Evaluación de una política: Teniendo una política cualquiera π, ¿cómo obtener la función V^π?
  Sabemos que la función V^π debe satisfacer la ecuación de Bellman:

  $$V^\pi(s) = \mathbb{E}[r + \gamma V^\pi(s')|s_t = s].$$

  Pero esto nos deja con el mismo problema, pues tenemos de ambos lados una función desconocida. Para problemas de decisión finitos, este es un sistema de ecuaciones con tantas ecuaciones como estados, lo cual puede ser costoso de resolver. Consideramos entonces un método iterativo:

  $$V_{k+1}(s) = \pi(s) \sum_{s'} \mathcal{P}^a_{s,s'}[\mathcal{R}^a_{s,s'} + \gamma V_k(s')],$$

  donde, $\mathcal{P}^a_{s,s'}$ es la probabilidad de pasar al estado s' desde s tomando la acción a. Y $\mathcal{R}^a_{s,s'}$ es el valor esperado de la siguiente recompensa al pasar de s a s' tomando la acción a.

- Mejora de una política: Dada una política π y estando en un estado s nos preguntamos qué pasaría si en vez de escoger la acción $\pi(s)$ escogieramos cualquier otra acción a. La consecuencia de hacer esto estaría dada por $Q^\pi(s,a)$. Entonces, la clave es comparar esto con V^π. De ser mayor que la función valor en ese estado, entonces estaríamos encontrando una mejor política, una que determina $\pi'(s) = a$ en vez de $\pi(s)$. Este es un caso particular de un resultado conocido como *Teorema de Mejora de Polítca* que dice que si π y π' son dos políticas tales que $Q^\pi(s, \pi'(s)) \geq V^\pi(s)$ entonces π' es mejor que π; esto es, $V^{\pi'}(s) \geq V^\pi(s)$ para todo estado s. Entonces, resulta natural querer buscar estas mejoras para dos los estados posibles y todas las acciones posibles y obtener aquellos que maximicen la función Q; es decir, queremos construir una nueva política π' como:

  $$\pi'(s) = \arg\max_a Q^\pi(s,a).$$

  Para proceso de decisión finitos, siempre encontramos una mejor política, a no ser que ya estemos en el óptimo.

- Iteración de políticas: Una vez que una policy π' ha sido obtenida de mejorar π utilizando V^π, podemos construir $V^{\pi'}$ y repetir nuevamente el proceso:

  $$\pi_0 \xrightarrow{E} V^{\pi_0} \xrightarrow{I} \pi_1 \xrightarrow{E} V^{\pi_1} \xrightarrow{E} \pi_2 \xrightarrow{E} \cdots \xrightarrow{I} \pi^* \xrightarrow{E} V^*$$

  pues cada política está garantizado que es mejor que la anterior.

- Iteración de valor: El método anterior puede ser muy tardado pues la convergencia, aunque garantizada teóricamente, sólo se da en el límite. Podemos truncar la etapa de policy evaluation.

  $$V_{k+1}(s) = \max_a \mathbb{E}[r + \gamma V_k(s')|s_t = s, a_t = a].$$

- Iteración de política generalizada (GPI): Notar que la iteración de políticas consiste en dos procesos que interactuan entre sí, pero de manera que cada proceso se hace de manera completa antes de comenzar el otro lleva a preguntarse si esto puede hacerse de manera más eficiente. Se conoce como iteración de políticas generalizada (GPI) a la idea general de que los procesos de evaluación y mejora de política interactuen entre sí.

- Limitaciones: Los métodos de programación dinámica suponen que se cuenta con un modelo completo del sistema, lo cual es prácticamente imposible en la práctica, además de ser muy costosos en términos computacionales. Estos métodos sufren de la llamada *maldición de la dimensionalidad*, pues el número de estados crece exponencialmente con el número de variables

7.2. Métodos básicos de solución: Model-free

Métodos que no necesitan un modelo completo del sistema, sólo necesitan experiencia. Encontramos los métodos Monte Carlo que simulan las transiciones y los estados, y los métodos de diferencias temporales, que van actualizando iterativamente el valor de los estados observados, el más famoso de estos es el TD(λ).

7.2.1. Métodos Monte-Carlo

Los métodos Monte-Carlo no requieren un modelo del ambiente y son conceptualmente sencillos, pero no son adecuados para hacer actualizaciones incrementales. Estos métodos requieren *experiencia* en vez de un modelo completo; es decir, requieren secuencias de estados, acciones y recompensas, ya sean estos obtenidos de simulaciones o de interacción *on-line* con el ambiente. Los métodos Monte-Carlo suponen que la experiencia proviene de eventos episódicos, es decir que terminan siempre de una u otra forma y es al final de estos episodios que se alteran las estimaciones de valor así como las políticas. En este sentido, los métodos Monte-Carlo son incrementales por episodio, pero no paso a paso dentro de cada episodio. A pesar de las diferencias operativas entre los métodos DP y los métodos monte-carlo, las ideas esenciales son las mismas, pues ambos métodos calculan las mismas funciones además de seguir el mismo camino hacia la optimalidad; es decir, evaluación, mejora y finalmente iteración general. La siguiente descripción de los métodos Monte-Carlo está basada en el clásico libro de **?**

- Evaluación Monte-Carlo: Como se mencionó antes, una manera de estimar el valor de un estado sería promediar las recompensas obtenidas al visitar un estado entre el número de visitas a este; por la ley de los grandes números este promedio debe converger al valor real. Dado un conjunto de episodios obtenidos de seguir una política π, denotamos como *visita* cada vez que se llega a un estado s. El algoritmo *every-visit* estima $V^\pi(s)$ como las recompensas promedio de todas las visitas al estado s en todos los episodios; por otro lado, el algoritmo *first-visit* promedia sólo lo obtenido al visitar por primera vez un estado en todos los episodios. Ambos métodos convergen a V^π.

- Notemos que al no tener un modelo del ambiente, es más útil estimar el valor de una acción más que el valor de un estado. Teniendo un modelo, los valores de los estados son suficientes para determinar una política, pues uno simplemente ve un paso hacia adelante y escoge la acción que tenga la mejor combinación de recompensa y de estado siguiente, pero sin un modelo disponible, uno debe estimar el valor de las acciones para con esto obtener una política. Entonces, es importante tener un método para estimar Q^*.

- Estimación de $Q^\pi(s, a)$: Los métodos son esencialmente los mismos que para estimar la función valor; es decir, promediando las recompensas obtenidas al visitar un estado y haber seleccionado cierta acción (todas las visitas) y análogamente para la primer visita. El problema que tienen estos métodos es que no todas las parejas acción-estado son visitadas, pues para obtener estimaciones válidas deberíamos tener parejas estado-acción para todas las acciones y no sólo las que se ven en los episodios.

- Control Monte-Carlo: para una primera versión Monte-Carlo de iteración de políticas, supondremos que contamos con un número infinito de episodios y que se visitan todas las parejas estado-acción dentro de

cada episodio. A diferencia de los métodos de DP y por las razones que ya se explicaron, el proceso de evaluación y mejora se hace respecto a la función $Q(s, a)$.

$$\pi_0 \xrightarrow{E} Q^{\pi_0} \xrightarrow{I} \pi_1 \xrightarrow{E} Q^{\pi_1} \xrightarrow{E} \pi_2 \xrightarrow{E} \cdots \xrightarrow{I} \pi^* \xrightarrow{E} Q^*.$$

Bajo el supuesto de episodios infinitos y visita a todas las parejas estado-acción se tiene que los métodos Monte-Carlo encuentran Q^{π_k} de manera exacta para π_k.
Para hacer la mejora de una política, se escoge de manera determinística la acción a en la cual se alcanza el máximo de la función Q; es decir,

$$\pi(s) = \arg\max_a Q(s, a).$$

- Para remover el supuesto de un número infinito de episodios, una opción es abandonar la idea de obtener completamente la función Q_k^π en el paso de evaluación antes de proceder a la mejora, sino sólo mover un poco la función Q actual hacia la verdadera.
- En cuanto al supuesto de visita a todas las parejas acción-estado la única alternativa a suponer que se visiten todas las parejas es hacer que el agente las seleccione y para esto existen dos familias de métodos:
  - On-Policy: intentan evaluar mejorar la política que es actualmente utilizada para la toma de decisiones.
  - Off-Policy: Evaluan o mejoran una política distinta

7.2.2. Diferencias Temporales

Los métodos TD son totalmente incrementales y tampoco requieren un modelo del ambiente, pero son mucho más complejos. Estos métodos son una combinación de métodos Monte-Carlo y de Programación Dinámica; estos métodos pueden aprender directamente de la experiencia y no necesitan un modelo del ambiente, pero también utilizan técnicas de *bootstrapping* (estiman usando estimadores) por lo que no necesitan finalizar completamente un episodio para actualizar los valores.

- Evaluación de política en Diferencias Temporales
- TD(0)
- SARSA: Es un método *on-policy*
- Q-Learning: Es un método *off-Policy*
- Métodos Actor-Critic

8. Deep Reinforcement Learning

8.1. Mejoras a los métodos básicos

- Programación Dinámica con aproximación de funciones.
- Cuando el espacio de estados es muy grande, no es factible utilizar una representación en una *look up table*, por lo que es necesario utilizar otro tipo de aproximadores (**?**)

8.2. Deep Learning

- Redes neuronales profundas como aproximadores generales.
- Redes Convolucionales para poder tomar como input los pixeles de un juego directamente.

8.3. Value based Reinforcement Learning:

- Representar la función valor mediante una *Q-network* con pesos w:

$$Q(s, a, w) \approx Q^*(s, a)$$

- Q-learning:
  - Sabemos que los valores óptimos deben satisfacer la ecuación de Bellman:

$$Q^*(s, a) = \mathbb{E}_{s'}[r + \gamma \max_{a'} Q^*(s', a')|s, a].$$

  - Tratamos el término $r + \gamma \max_{a'} Q^*(s', a', w)$ como el objetivo.
  - Minimizar error cuadrático medio mediante Gradiente Estocástico:

$$I = (r + \gamma \max_{a'} Q^*(s', a', w) - Q(s, a, w))^2.$$

  - Converge a Q^* utilizando una representación tabular.
  - Diverge al utilizar redes neuronales debido a la alta correlación entre muestras y la no-estacionareidad del objetivo.
  - Lo resolvemos con Experience Replay (**?**), el cual consiste en guardar todas las experiencias de modo que al entrenar la red tomar sólo mini-batches aleatorios en vez de utilizar la transición más reciente.
- Aplicaciones de DQN.
  - **?**. Paper en el que muestran cómo una computadora aprende a jugar Atari 2600 a través de sólo observar los pixeles del juego y obteniendo una recompensa conforme el score aumentaba. Utilizando la misma arquitectura, se pudieron aprender 7 juegos
  - **?**. En este paper, aumentaron a 49 juegos aprendidos. Utilizan una Deep Q-Network, que consiste en aproximar la función Q mediante una red neuronal que toma como input los estados (las pantallas del juego) y devuelve como output el valor Q para cada posible acción. En este caso, utilizaron una red convolucional con 3 capas de convolución y 2 capas totalmente conectadas.
- Mejoras posteriores a DQN
  - Double DQN.
  - Prioritized Replay.
  - Duelling Networks.

8.4. Policy Based DRL

Resulta que Deep Q-Learning no es un gran algoritmo, y en su lugar, **?** proponen utilizar directamente Policy-Gradients. La idea es representar una policy mediante una red neuronal profunda con pesos $\mathbf{u}$.

$$a = \pi(a|s, u) \text{ o } a = \pi(s, u).$$

En este contexto, definimos una función objetivo $L(\mathbf{u})$ como:

$$L(\mathbf{u}) = \mathbb{E}[r_1 + \gamma r_2 + \gamma^2 r_3 + ...|\pi(\cdot, \mathbf{u})].$$

Optimizamos con Stochastic Gradient Descent. El gradiente de una policy estocástica está dado $\pi(a|s, \mathbf{u})$ por:

$$\frac{\partial L(\mathbf{u})}{\partial \mathbf{u}} = \mathbb{E}\left[\frac{\partial \log \pi(a|s, \mathbf{u})}{\partial \mathbf{u}} Q^\pi(s, a)\right].$$

Por otro lado, el gradiente de una policy determinística $a = \pi(s)$ está dado por:

$$\frac{\partial L(\mathbf{u})}{\partial \mathbf{u}} = \mathbb{E}\left[\frac{\partial Q^{\pi}(s,a)}{\partial a}\frac{\partial a}{\partial \mathbf{u}}\right].$$

Los métodos *actor-only*, donde *actor* se utiliza como sinónimo de *policy* utilizan familias parametrizadas de policies. Las ventajas de tener una policy parametrizada es que podemos generar un espectro continuo de accciones, pero los métodos de optimización (policy gradients) sufren de alta varianza en sus estimaciones del gradiente, lo cual genera un lento aprendizaje (**?**,**?**,**?**,**?**,**?**,**?**).

Los métodos *critic-only*, donde *critic* se utiliza como sinónimo de *value function*, que utilizan diferencias temporales tienen una menor varianza en las estimaciones del retorno esperado (**?**,**?**,**?**). Una manera de derivar una policy en estos escenarios es seleccionando greedy actions; es decir, acciones praa las cuales la value funtion indique que el retorno esperado es máximo, lo cual puede requerir un alto costo computacional. Lo usual es discretizar el espacio continuo de acciones. Los métodos *actor-critic*, combinan las vetajas de ambos mundos. Mientras que las familias parametrizadas tienen la ventaja de utilizar acciones continuas, los métodos critic-only proveen de conocimiento de baja varianza sobre el desempeño; es decir, la estimación del retorno esperado del crític-only permite al actor-only actualizar los gradientes con información de menor varianza, lo cual acelera todo el proceso. La menor varianza es intercambiada por un mayor sesgo al inicio del aprendizaje (**?**).

Los métodos de Policy Gradients optimizan directamente la recompensa acumulada y pueden utilizarse junto con aproximadores no-lineales como las redes neuronales. Este método presenta dos grandes retos: el número de muestras de entrenamiento requeridos y la dificultad de obtener mejoras estables debido a la no-estacionariedad de los datos. **?** atacan el primer problema se puede atacar utilizando value functions para reducir la varianza de los estimados. El segundo problema se ataca utilizando un método de región de confianza tanto para la política como para la value function, los cuales están representados por redes neuronales.

? Proponen "Trust Region Policy Optimization", Procedimiento iterativo para optimizar políticas con mejora monótona garantizada. Es un algoritmo similar a Policy Gradients y resulta útil para optimizar políticas con redes neuronales. Buen performance en juegos de Atari que utilizan la pantalla como input.

9. Reinforcement Learning y Causalidad

Aprendemos causalidad mediante intervenciones, interacciones, etc. Reinforcement Learning es aprender a realizar tareas mediante interacción. ¿Habrá algo en medio?

9.1. Conocimiento causal y observabilidad parcial

Los algoritmos actuales de reinforcement Learning suponen que el problema puede modelarse dentro de un contexto Markoviano de decisión con recompensas. Si el supuesto markoviano se elimina, los algoritmos podrían dejar no funcionar, como muestran en **?**. Además, todo el desarrollo teórico utiliza la propiedad Markoviana.

Una manera en la que la propiedad Markoviana puede perderse es si el agente no cuenta con información completa sobre el estado; es decir, al igual que en los HMM, suponemos que el ambiente es Markoviano, pero el agente no tiene acceso a él de modo que los datos no le *parecen* Markovianos. A diferencia de los HMM, no existe un procedimiento computacionalmente tratable para los POMDP. El problema es que una vez que se obtienen las estimaciones del valor para cada estado, se debe hacer programación dinámica en espacios continuos y esto es infactible salvo para casos pequeños

La manera de generalizar el caso markoviano al caso no observado es considerar un MDP subyacente con estados $\mathcal{S} = \{s_1, ..., s_N\}$ y probabilidades de transición $\mathcal{P}^a_{s,s'}$. Luego, en cada tiempo t, el agente observa un mensaje m_t el cual proviene de una distribución $P(m|s_t)$ desconocida.

En **?** se describe un procedimiento que evita estimar el valor de los estados y trabaja directamente en el espacio de políticas estocásticas. El procedimiento es *directo*, a diferencia de los métodos *indirectos* en los cuales primero se aprenden los parámetros del MDP y luego se utiliza programación dinámica para obtener la política. El algoritmo propuesto consiste en dos partes: primero, se utiliza un método Monte-Carlo para calcular un análogo de los Q-values y luego se utiliza un paso de policy-improvement.

Desde un punto de vista de psicología, una manear de entender la inferencia de estados ocultos es en términos de causas latentes (**?**). Siguiendo esta idea, es evidente que un agente de RL que opera en el mundo real no sólo debe hacer inferencia de estados ocultos, sino que debe poder descubrir los estados ocultos que están detrás del mundo que éste observa. Esto es una forma de aprendizaje estructural latente (**?**).

Un primer acercamiento a este problema puede hacerse mediante técnicas de inferencia Bayesiana no paramétrica. Por ejemplo, tanto **?** como **?** utilizan priors no paramétricas para hacer inferencia sobre causas latentes. Un modelo causal latente es una buena representación de la estructura causal verdadera bajo condicionamiento Pavloviano, pues es el experimentador quien es una causa latente.

En líneas más avanzadas, **?** exploraron versiones que permiten que múltiples causas latentes estén activas.

9.2. Towards Deep Symbolic Reinforcement Learning

Deep Reinforcement learning consiste, básicamente, en utilizar una red profunda para aproximar la Q-function; por lo tanto, los métodos de DRL heredan los problemas del Deep Learning: se requiere un gran volumen de datos, son muy lentas en entrenarse y, según **?**, son esencialmente *reactivos*, lo que significa que no utilizan procedimientos de alto nivel como planeación o razonamiento causal, además de que no son interpretables.

? proponen una arquitectura híbrida que, combina redes neuronales con inteligencia artificial simbólica. El modelo propuesto consiste en combinar una red neuronal que como entrada directamente lo que se recibe del ambiente y transforma estos datos crudos en una representación simbólica basada en lógica de primer orden; esta representación simbólica es posteriormente utilizada para escoger la acción a tomar.

Se utiliza un sencillo juego para probar la arquitectura y en este contexto la red neuronal debe generar representaciones simbólicas de secuencias de estados en el juego en las cuales el flujo de pixeles queda codificado en términos de objetos, tipos y sus interaccioes. Luego, con esto se aprenden las políticas.

El primer paso que hacen es utilizar un autoencoder convolucional para generar, de manera no supervisada, un conjunto de símbolos que representarán los objetos en cada escena. Luego, para cada pixel, obtienen la feature con mayor activación y se toman los más altos. De esta manera se obtiene un pixel representativo para cada objeto a los cuales posteriormente se le asigna un tipo simbólico según el autoencoder.

Luego, una vez que se tienen estos símbolos de bajo nivel para cada cuadro del juego, se necesita rastrearlos a lo largo del tiempo. Utilizan una verosimilitud que es función de proximidad espacial y el tipo de transiciones. Para considerar la información de la dinámica de los cuadros utilizan la diferencia entre cuadros y las posiciones relativas entre objetos.

Para la parte de aprendizaje, entrenan una función Q por separado para cada interacción entre objetos con actualización

$$Q^{ij}(s_t^{ij}, a_t) \leftarrow Q^{ij}(s_t^{ij}, a_t) + \alpha \left(r_{t+1} + \gamma(\max_a Q^{ij}(s_{t+1}^{ij}, a) - Q^{ij}(s_t^{ij}, a_t)) \right).$$

Se prueba el framework sobre varias variantes de un sencillo videojuego.

Llama la atención como estos autores utilizan maquinaria de lo más reciente (deep learning, autoencoders)

para luego obtener una representación en términos de lógica de primer orden. Este approach ha quedado claro que ya quedó atrás por lo que debería buscarse algo que sustituya este enfoque lógico. En particular, pudiera ser esto un modelo gráfico causal el cual vaya aprendiéndose conforme el agente de RL aprende sobre su entorno.

9.3. Conocimiento causal y toma de decisiones secuenciales: acciones como intervenciones

En cuanto al aprendizaje *on-line* de modelos causales, encontramos en los trabajos de **?** y de **?** acercamientos tangenciales, o un punto de inicio, al aprendizaje de modelos causales al vuelo.

Teorías de cómo las personas usan datos de covariación para aprender modelos gráfios causales pueden dividirse en enfoques racionales y heurísticos (según **?**). El enfoque racional modela el aprendizaje causal como inferencia racional, lo que incluye la inferencia Bayesiana. Estos modelos se mantienen en el nivel computacional y no intentan explicar el mecanismo cognitivo subyacente.

Por otro lado, el enfoque heurístico propone que las personas reaccionan a distintos estímulos que los hacen moficiar sus hipótesis causales de una manera no necesariamente racional. El modelo de efecto simple de **?** supone que las personas se enfocan en evaluar una sola relación causal a la vez. Es un modelo que aprende la magnitud de la relación causal. Si una variable supera cierto umbral, entonces la persona acepta esa relación causal y la incorpora a su bagaje de relaciones.

El modelo propuesto por **?** sigue esta línea del efecto simple, con la diferencia de que este modelo sí intenta descubrir el proceso cognitivo subyacente, pues el aprender magnitudes causales es un proceso puramente computacional. El modelo de efecto simple puede verse como una implementación del modelo de **?**. Además, según narran **?**, el modelo de efecto simple no explica cómo la persona utiliza las magnitudes causales para determinar si existe o no una relación causal. En el modelo propuesto, **?**, incorporan un procedimiento de decisión para determinar si existen o no relaciones causales.

El modelo LPL (**?**) comienza con una estructura causal inicial, la cual consiste en un grafo acíclico dirigido el cual representa las creencias iniciales de un individuo. Posteriormente, el modelo altera esta estructura al añadir o remover una sola arista, de manera que el problema se reduce a evaluar relaciones causales entre dos variables. Este modelo genera estimaciones de la fuerza de la relación causal para cada posible relación causa-efecto que no sea eliminada a priori.

Supongamos que un individuo cree a priori que B causa E, y está tratando de averiguar si C causa E. A diferencia del modelo de efecto simple, las estimaciones de la fuerza causal están generadas por un mecanismo similar a los filtros de partículas Bayesianos En particular, a partir del conocimiento causal inicial del individuo, el modelo LPL comienza con n partículas. Estas partículas $\{V_C^1, ..., V_C^n\}$ representan las creencias actuales del individuo sobre si C causa E. Paralelamente, existen $\{V_B^1, ...V_B^n\}$ para B. Se define la *capa* i de partículas para el efecto E como la i-ésima partícula de cada causa conocida y posible para E. En el ejemplo, tendríamos que esta capa es $\{V_B^i, V_C^i\}$. Luego, el modelo LPL actualiza cada capa de manera independiente de la siguiente manera

$$V_C^{i+1} = V_C^{i,t} + \alpha(\text{observado} - \text{esperado}),$$

Donde "observado" tiene el valor de 1 si el efecto sucede y 0 en caso contrario. Por otro lado, "esperado" es el valor esperado del efecto, calculado según la forma funcional de la relación causa-efecto, el cual es calculado por separado para cada causa potencial en una capa. El valor esperado de E (para la capa i y la causa potencial C) es

$$\text{esperado} = \prod_k (1 - V_k^{t-1})(1 - \prod_j (1 - V_j^{t-1}).$$

Conforme el individuo observa más datos, las partículas convergen a los valores verdaderos de los parámetros y las estimaciones para cada capa convergen. Este procedimiento no toma en cuenta aspectos estructurales, los cuales se modifican según un criterio de decisión. Lo que hace el modelo es una prueba t en la cual la hipótesis nula es que las partículas provienen de una distribución con media cero. El criterio de decisión es el siguiente: si

actualmente no existe una arista, y la prueba t rechaza la hipótesis nula, entonces se añade la arista. Si existe una arista, y la prueba no es significativa, entonces se elimina la arista existente. Alterar una arista afectará las futuras estimaciones del valor "esperado" para otras causas.

? llaman la atención sobre que los modelos actuales de Decisiones (maximum expected utility) no asumen que los tomadores de decisiones aprenden sobre la estructura causal del problema. Por otro lado, las teorías de aprendizaje causal proponen que un tomador de decisiones guían sus acciones a través de conocimiento causal previamente obtenido. El modelo de **?** extiende las teorías de aprendizaje causal al contexto de toma de decisiones, la idea básica consiste en que los tomadores de decisiones utilizan la información disponible de su entorno para inducir un modelo causal del problema de decisión al que se enfrentan. Se sabe que en los seres humanos el conocimiento causal guía la interpretación de información probabilista y que los humanos somos capaces de aprender la estructura causal del entorno a través de una serie de (traducir cues) (**?**). La pregunta fundamental que plantean **?** es entender qué es lo que aprenden las personas al tomar decisiones de manera secuencial de tal forma que se busque maximizar cierta recompensa a largo plazo y donde las acciones alteran el entorno. Al exponerse una persona a estas situaciones de decisión secuencial se obtiene una retroalimentación de cómo las acciones se relacionaban con la obtención de recompensas. Las distintas teorías sobre toma de decisiones bajo incertidumbre suponen distintas representaciones para esta retroalimentación. Hagmayer propone agrupar estas distintas representaciones en tres niveles:

- Nivel 1: Opciones, resultados y recompensas. Sólo los valores esperados de las opciones posibles afectan la toma de decisiones. Se conoce como *habit learning* en la literatura de aprendizaje en animales.
- Nivel 2: Opciones, relaciones causales de estas opciones a los resultados, relaciones causales entre variables de resultado y relaciones de estas con las recompensas. Aquí entra lo que se conoce como *goal-directed learning*. La diferencia con el nivel 1 es la capacidad de distinguir entre acciones y los valores asociados a estas.
- Nivel 3: Opciones, relaciones causales entre variables de resultado, relaciones causales entre estas variables y relaciones de estas con las recompensas. Estos son los modelos causales. Las distintas teorías sobre estos modelos suponen que los humanos (o el agente que aprende) adquieren estos modelos mediante procesos de aprendizaje. Algunas teorías suponen que la estructura causal del entorno se obtiene a través de *claves*, como la propuesta por **?**. Por otro lado, otros autores opinan que las personas utilizan propiedades estadísticas de los datos observados, este sería el caso de las teorías de **?** y de **?**.

Aunque las teorías que entran en el nivel 2 son sensibles a las correlaciones entre las respuestas obtenidas, sólo en el nivel 3 se hace una distinción entre causalidad y correlación.

Los autores hacen énfasis en que aunque las personas sí utilicen modelos causales al tomar decisiones simples, estas decisiones suelen ser sobre escenarios hipotéticos, pero no se han hecho muchos estudios sobre adquisición de conocimiento causal al enfrentarse a problemas de decisión secuencial con el objetivo de maximizar alguna ganancia a largo plazo. Además, en los estudios realizados, se ha visto que las personas tienen la capacidad de inducir modelos causales a partir de claves en el entorno; se sabe además que a partir de intervenciones activas se toman mejores decisiones que considerando sólo datos observados de manera pasiva y que las personas tienen la capacidad de integrar relaciones causales obtenidas de manera separada. Pero en estos estudios las personas tenían como objetivo descubrir de manera directa estructuras causales. No se sabe bien hasta qué punto las personas adquieren conocimiento causal al perseguir objetivos que no sean descubrir directamente estructuras causales a partir del ambiente.

Los autores realizaron tres experimentos para explorar las representaciones que las personas hacen de las relaciones causales que adquieren al tomar decisiones para maximizar algún objetivo. En el primer experimento, el objetivo era estudiar cómo las creencias causales de las personas alteran la toma de decisiones y cómo los participantes utilizaban los resultados de sus decisiones para revisar el modelo causal. El experimento consistía en dos etapas de toma de decisiones, la segunda siendo una etapa de prueba. En ambas etapas los participantes

debían escoger entre una serie de opciones de modo que maximizaran ganancias. Se le sugirió a cada participante una posible cadena causal. En la primera etapa se les proveía retroalimentación de las consecuencias de sus decisiones, lo cual no sucedió en la segunda etapa sino que ellos mismos debían inferir posibles relaciones causales. Se concluyó que los participantes encontraron estructura causal en el problema y la utilizaron para tomar futuras decisiones.

En el segundo experimento y tercer experimento, a los participantes se les mostró una serie de hipótesis causales para ver si inducían de manera espontánea un modelo causal de la situación.

Lo que estos experimentos muestran, en conclusión de los autores, es que al intentar maximizar una ganancia a largo plazo, las personas adquieren conocimiento causal del entorno, el cual afecta futuras decisiones. Mencionan que estos experimentos ponen en tela de juicio los modelos estándar de toma de decisiones, pues utilizando sólo valores esperados no se llegarían a los mismos resultados.

Como consecuencia de sus estudios, se obtiene que los métodos basados en los niveles 1 y 2, que es donde caerían los métodos tradicionales de aprendizaje por refuerzo no son suficientes para una adecuada toma de decisiones en un sistema causal, pues estas formulaciones no cuentan con la expresividad suficiente para capturar las relaciones causales. Esto confirma la necesidad de pensar en modelos que permitan aprender modelos causales a la par que un agente aprende de su entorno mediante interacción y exploración.

En cuanto a la formulación de un problema de decisión en sistemas causales, ya desde **?** tenemos una idea de cómo aumentar redes bayesianas con nodos de decisión y variables de estrategia para el modelado causal en problemas de decisión.

? plantean que cada decisión tomada por un agente es una intervención directa con el fin de predecir las consecuencias de esta.

10. Propuesta

10.1. Pregunta de Investigación

¿Los métodos de aprendizaje por refuerzo pueden aprender un modelo causal del ambiente? Más allá de producir un mapeo estado-acción, queremos obtener un modelo causal del entorno.

10.2. Problema

10.3. Objetivos

- Aprendizaje causal en línea a partir de acciones y recompensas obtenidas por un agente.
- Evaluación del modelo causal.
- Incorporar el conocimiento causal a la toma de decisiones en un loop de mejora continua.
- Exportar conocimiento causal obtenido para hacer inferencia causal.

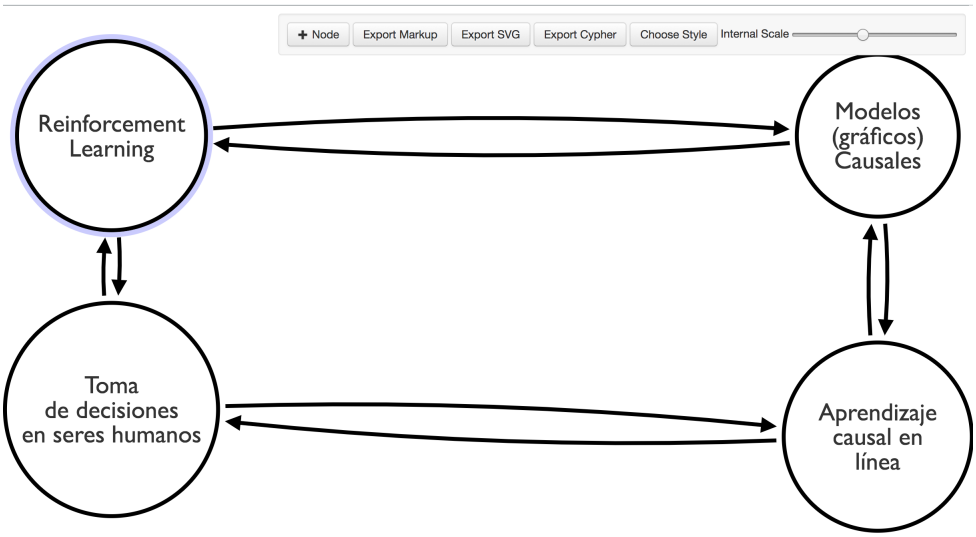
10.4. Propuesta metodológica

Dado el contexto proporcionado en las secciones anteriores, la propuesta consiste en incorporar un modelo de aprendizaje causal dentro a un problema de toma de decisiones. En el contexto de aprendizaje por refuerzo, tendremos un agente que, al tener como objetivo aprender una tarea mediante interacciones secuenciales, aprenda con cada paso un poco sobre la *estructura causal* de su entorno. Cada decisión que el agente toma (o acción que lleva a cabo) puede verse como una intervención en el ambiente y que proporciona información sobre la estructura causal de este. Esta información obtenida, será utilizada para mejorar las decisiones que toma y

para aprender otros aspectos estructurales (causales) sobre su ambiente. La intuición detrás de esto es que al aprender mediante una toma de decisiones repetida, decisiones que intervienen directamente sobre el entorno, se aprenden también relaciones causales; por ejemplo, al aprender a conducir se aprende implícitamente que ciertos movimientos del volante *causan* ciertos cambios en el estado del automovil (**?**). Otro ejemplo es en el contexto de videojuegos, en el cual al tomar ciertas decisiones se conoce el impacto que estas provocan en el estado del juego, conocimiento que a su vez es utilizado para tomar mejores decisiones en un futuro y que maximicen las recompensas u objetivos a largo plazo. No sólo hay intuición detrás de esto, pues se sabe que los seres humanos utilizan el conocimiento causal que la experiencia les provee para realizar predicciones para intervenciones en el mundo no realizadas aun (**?**, **?**)

El trabajo de **?** nos muestra que efectivamente pueden aprenderse aspectos estructurales sobre el entorno mientras el agente aprende, incluso logrando así mejorar su desempeño. Además, muestra la importancia de tener una semántica que permita hablar de la estructura del entorno. Ellos utilizaron lógica de primer orden, pero un modelo gráfico sería más expresivo y adaptable.

A su vez, el trabajo de **?** muestra cómo un modelo gráfico causal podría ser aprendido en línea al tomar en cuenta las decisiones tomadas y las recompensas obtenidas por un tomador de decisiones. Por otro lado, los modelos de **?**, **?**, **?** y **?** muestran que estas intuiciones no son erradas y que los seres humanos incorporamos y adquirimos conocimiento causal al enfrentarnos en problemas de toma de decisión simples, aunque aun no existen muchos estudios en los cuales los escenarios de decisión no sean hipotéticos sino que se intente maximizar una recompensa real a largo plazo. **?** mencionan que faltan estudios en donde las personas tengan otro objetivo más allá de revelar directamente la estructura causal del entorno y en los cuales vean directamente las consecuencias de sus acciones. Entonces, queda abierta la pregunta de qué pasaría con un agente que se enfrente a un problema de aprendizaje por refuerzo, queremos saber cómo este agente aprenderá paso a paso un modelo causal y lo utilizará para tomar mejores decisiones.



Referencias

Ballard, D. H. (2015). *Brain computation as hierarchical abstraction*. MIT Press.

Baxter, J. and Bartlett, P. L. (2001). Infinite-horizon policy-gradient estimation. *Journal of Artificial Intelligence Research*, 15:319–350.

Berenji, H. R. and Vengerov, D. (2003). A convergent actor-critic-based frl algorithm with application to power management of wireless transmitters. *IEEE Transactions on Fuzzy Systems*, 11(4):478–485.

Bermúdez, J. L. (2014). *Cognitive science: An introduction to the science of the mind*. Cambridge University Press.

Bertsekas, D. P., Bertsekas, D. P., Bertsekas, D. P., and Bertsekas, D. P. (1995). *Dynamic programming and optimal control*, volume 1. Athena Scientific Belmont, MA.

Bojduj, B. N. (2009). Extraction of causal-association networks from unstructured text data. *Master's Theses and Project Reports*, page 138.

Bonawitz, E., Denison, S., Chen, A., Gopnik, A., and Griffiths, T. (2011). A simple sequential algorithm for approximating bayesian inference. In *Proceedings of the Cognitive Science Society*, volume 33.

Boyan, J. A. (2002). Technical update: Least-squares temporal difference learning. *Machine Learning*, 49(2-3):233–246.

Brown, H., Friston, K. J., and Bestmann, S. (2011). Active inference, attention, and motor preparation. *Frontiers in psychology*, 2:218.

Bunescu, R. and Mooney, R. J. (2007). Statistical relational learning for natural language information extraction. *Introduction to Statistical relational learning*, pages 535–552.

Busoniu, L., Babuska, R., De Schutter, B., and Ernst, D. (2010). *Reinforcement learning and dynamic programming using function approximators*, volume 39. CRC press.

Cassandra, A. R., Kaelbling, L. P., and Littman, M. L. (1994). Acting optimally in partially observable stochastic domains. In *AAAI*, volume 94, pages 1023–1028.

Chickering, D. M. (2002). Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov):507–554.

Clark, A. (2013a). Expecting the world: perception, prediction, and the origins of human knowledge. *The Journal of Philosophy*, 110(9):469–496.

Clark, A. (2013b). The many faces of precision (replies to commentaries on "whatever next? neural prediction, situated agents, and the future of cognitive science"). *Frontiers in psychology*, 4:270.

Clark, A. (2013c). Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36(03):181–204.

Clark, A. (2014). Perceiving as predicting. *Perception and its modalities*.

Clark, A. (2015a). Embodied prediction. In *Open MIND*. Open MIND. Frankfurt am Main: MIND Group.

Clark, A. (2015b). *Surfing uncertainty: Prediction, action, and the embodied mind*. Oxford University Press.

Colombo, D., Maathuis, M. H., Kalisch, M., and Richardson, T. S. (2012). Learning high-dimensional directed acyclic graphs with latent and selection variables. *The Annals of Statistics*, pages 294–321.

Courville, A. C., Daw, N. D., and Touretzky, D. S. (2006). Bayesian theories of conditioning in a changing world. *Trends in cognitive sciences*, 10(7):294–300.

Culotta, A., McCallum, A., and Betz, J. (2006). Integrating probabilistic extraction models and data mining to discover relations and patterns in text. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 296–303. Association for Computational Linguistics.

Danks, D. (2014). *Unifying the mind: Cognitive representations as graphical models*. Mit Press.

Danks, D., Griffiths, T. L., and Tenenbaum, J. B. (2003). Dynamical causal learning. In *Advances in neural information processing systems*, pages 83–90.

Dawid, A. P. (2002). Influence diagrams for causal modelling and inference. *International Statistical Review*, 70(2):161–189.

Dayan, P. and Abbott, L. F. (2001). *Theoretical neuroscience*, volume 10. Cambridge, MA: MIT Press.

Dayan, P. and Hinton, G. E. (1996). Varieties of helmholtz machine. *Neural Networks*, 9(8):1385–1403.

Dayan, P., Hinton, G. E., Neal, R. M., and Zemel, R. S. (1995). The helmholtz machine. *Neural computation*, 7(5):889–904.

Fox, C. W., Girdhar, N., and Gurney, K. N. (2008). A causal bayesian network view of reinforcement learning. In *FLAIRS Conference*, pages 109–110.

Friston, K. (2005). A theory of cortical responses. *Philosophical transactions of the Royal Society B: Biological sciences*, 360(1456):815–836.

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 11(2):127–138.

Garnelo, M., Arulkumaran, K., and Shanahan, M. (2016). Towards deep symbolic reinforcement learning. *arXiv preprint arXiv:1609.05518*.

Gershman, S. J. (2015). Reinforcement learning and causal models. In *The Oxford Handbook of Causal Reasoning*.

Gershman, S. J. and Niv, Y. (2010). Learning latent structure: carving nature at its joints. *Current opinion in neurobiology*, 20(2):251–256.

Gershman, S. J. and Niv, Y. (2012). Exploring a latent cause theory of classical conditioning. *Learning & behavior*, 40(3):255–268.

Gershman, S. J., Norman, K. A., and Niv, Y. (2015). Discovering latent causes in reinforcement learning. *Current Opinion in Behavioral Sciences*, 5:43–50.

Girju, R. (2003). Automatic detection of causal relations for question answering. In *Proceedings of the ACL 2003 workshop on Multilingual summarization and question answering-Volume 12*, pages 76–83. Association for Computational Linguistics.

Girju, R., Moldovan, D. I., et al. (2002). Text mining for causal relations. In *FLAIRS Conference*, pages 360–364.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press. `http://www.deeplearningbook.org`.

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., and Danks, D. (2004). A theory of causal learning in children: causal maps and bayes nets. *Psychological review*, 111(1):3.

Goudet, O., Kalainathan, D., Caillou, P., Lopez-Paz, D., Guyon, I., Sebag, M., Tritas, A., and Tubaro, P. (2017). Learning functional causal models with generative neural networks. *arXiv preprint arXiv:1709.05321*.

Griffiths, T. L. and Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive psychology*, 51(4):334–384.

Griffiths, T. L. and Tenenbaum, J. B. (2009). Theory-based causal induction. *Psychological review*, 116(4):661.

Grondman, I., Busoniu, L., Lopes, G. A., and Babuska, R. (2012). A survey of actor-critic reinforcement learning: Standard and natural policy gradients. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(6):1291–1307.

Guyon, I. (2013). Chalearn cause effect pairs challenge. Technical report, •.

Hagmayer, Y. and Mayrhofer, R. (2013). Hierarchical bayesian models as formal models of causal reasoning. *Argument & Computation*, 4(1):36–45.

Hagmayer, Y. and Meder, B. (2008). Causal learning through repeated decision making. In *Proceedings of the Cognitive Science Society*, volume 30.

Hagmayer, Y. and Meder, B. (2013). Repeated causal decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 39(1):33.

Hagmayer, Y., Meder, B., Osman, M., Mangold, S., and Lagnado, D. (2010). Spontaneous causal learning while controlling a dynamic system. *The Open Psychology Journal*, 3:145–162.

Hagmayer, Y. and Sloman, S. A. (2009). Decision makers conceive of their choices as interventions. *Journal of Experimental Psychology: General*, 138(1):22.

Heckerman, D., Meek, C., and Koller, D. (2004). Probabilistic models for relational data. Technical report, Technical Report MSR-TR-2004-30, Microsoft Research.

Helmholtz, H. v. (1860). 1962. *Handbuch der physiologischen optik*, 3.

Hinton, G. E., Dayan, P., Frey, B. J., and Neal, R. M. (1995). The "wake-sleep" algorithm for unsupervised neural networks. *Science*, 268(5214):1158.

Hinton, G. E. and Zemel, R. S. (1994). Autoencoders, minimum description length, and helmholtz free energy. *Advances in neural information processing systems*, pages 3–3.

Hohwy, J. (2013). *The predictive mind*. Oxford University Press.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960.

Jaakkola, T., Singh, S. P., and Jordan, M. I. (1995). Reinforcement learning algorithm for partially observable markov decision problems. In *Advances in neural information processing systems*, pages 345–352.

Joyce, J. M. (1999). *The Foundations of Causal Decision Theory*. Cambridge University Press.

Karagas, M. R., Tosteson, T. D., Blum, J., Morris, J. S., Baron, J. A., and Klaue, B. (1998). Design of an epidemiologic study of drinking water arsenic exposure and skin and bladder cancer risk in a us population. *Environmental health perspectives*, 106(Suppl 4):1047.

Karagas, M. R., Tosteson, T. D., Morris, J. S., Demidenko, E., Mott, L. A., Heaney, J., and Schned, A. (2004). Incidence of transitional cell carcinoma of the bladder and arsenic exposure in new hampshire. *Cancer Causes and Control*, 15(5):465–472.

Kemp, C., Tenenbaum, J. B., Griffiths, T. L., Yamada, T., and Ueda, N. (2006). Learning systems of concepts with an infinite relational model. In *AAAI*, volume 3, page 5.

Khoo, C. S., Chan, S., and Niu, Y. (2000). Extracting causal knowledge from a medical database using graphical patterns. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pages 336–343. Association for Computational Linguistics.

Koller, D. and Friedman, N. (2009). *Probabilistic graphical models: principles and techniques*. MIT press.

Konda, V. R. and Tsitsiklis, J. N. (2003). Onactor-critic algorithms. *SIAM journal on Control and Optimization*, 42(4):1143–1166.

Kummerfeld, E. and Danks, D. (2013). Tracking time-varying graphical structure. In *Advances in neural information processing systems*, pages 1205–1213.

Lagnado, D. A., Waldmann, M. R., Hagmayer, Y., and Sloman, S. A. (2007). Beyond covariation. *Causal learning: Psychology, philosophy, and computation*, pages 154–172.

Lee, S. and Honavar, V. (2015). Lifted representation of relational causal models revisited: Implications for reasoning and structure learning. In *Proceedings of the UAI 2015 Conference on Advances in Causal Inference-Volume 1504*, pages 56–65. CEUR-WS. org.

Lee, S. and Honavar, V. (2016). On learning causal models from relational data. In *AAAI*, pages 3263–3270.

Lee, T. S. and Mumford, D. (2003). Hierarchical bayesian inference in the visual cortex. *JOSA A*, 20(7):1434–1448.

Lin, L.-J. (1992). Self-improving reactive agents based on reinforcement learning, planning and teaching. *Machine Learning*, 8(3):293–321.

Lopez-Paz, D., Muandet, K., Schölkopf, B., and Tolstikhin, I. (2015). Towards a learning theory of cause-effect inference. In *International Conference on Machine Learning*, pages 1452–1461.

Lucas, C. G. and Griffiths, T. L. (2010). Learning the form of causal relationships using hierarchical bayesian models. *Cognitive Science*, 34(1):113–147.

Maier, M., Marazopoulou, K., Arbour, D., and Jensen, D. (2013a). A sound and complete algorithm for learning causal models from relational data. *arXiv preprint arXiv:1309.6843*.

Maier, M., Marazopoulou, K., and Jensen, D. (2013b). Reasoning about independence in probabilistic models of relational data. *arXiv preprint arXiv:1302.4381*.

Marr, D. (1982). Vision: A computational investigation into the human representation and processing of visual information.

Meder, B., Hagmayer, Y., and Waldmann, M. R. (2008). Inferring interventional predictions from observational learning data. *Psychonomic Bulletin & Review*, 15(1):75–80.

Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T. P., Harley, T., Silver, D., and Kavukcuoglu, K. (2016). Asynchronous methods for deep reinforcement learning. *CoRR*, abs/1602.01783.

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., and Riedmiller, M. (2013). Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. (2015). Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533.

Mooij, J. M., Peters, J., Janzing, D., Zscheischler, J., and Schölkopf, B. (2016). Distinguishing cause from effect using observational data: methods and benchmarks. *The Journal of Machine Learning Research*, 17(1):1103–1204.

Pawar, S., Bhattacharyya, P., and Palshikar, G. (2016). End-to-end relation extraction using markov logic networks. In *Proceedings of the 17th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2016), LNCS*, volume 9624.

Pearl, J. (2009). *Causality*. Cambridge university press.

Pearl, J. (2014). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann.

Perrin, T., Kawai, H., Kunieda, K., and Yamada, K. (2008). Global dynamics network construction from the web. In *Information-Explosion and Next Generation Search, 2008. INGS'08. International Workshop on*, pages 69–76. IEEE.

Peters, J., Mooij, J., Janzing, D., and Schölkopf, B. (2012). Identifiability of causal graphs using functional models. *arXiv preprint arXiv:1202.3757*.

Puterman, M. L. (1994). *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., New York, NY, USA, 1st edition.

Quinn, J., Mooij, J., Heskes, T., and Biehl, M. (2011). Learning of causal relations. In *(ed.), ESANN 2011: Proceedings of the 19th European Symposium on Artificial Neural Networks*, pages 287–296. [Sl: sn].

Ramani, A. K., Bunescu, R. C., Mooney, R. J., and Marcotte, E. M. (2005). Consolidating the set of known human protein-protein interactions in preparation for large-scale mapping of the human interactome. *Genome Biology*, 6(5):R40.

Rasmussen, C. E. (2000). The infinite gaussian mixture model. In *Advances in neural information processing systems*, pages 554–560.

Richter, S., Aberdeen, D., Yu, J., et al. (2007). Natural actor-critic for road traffic optimisation. *Advances in neural information processing systems*, 19:1169.

Rink, B., Harabagiu, S., and Roberts, K. (2011). Automatic extraction of relations between medical concepts in clinical texts. *Journal of the American Medical Informatics Association*, 18(5):594–600.

Saito, H., Kawai, H., Tsuchida, M., Mizuguchi, H., and Kusui, D. (2007). Extraction of statistical terms and co-occurrence networks from newspapers. In *NTCIR*.

Sakai, H. and Masuyama, S. (2008). Cause information extraction from financial articles concerning business performance. *IEICE TRANSACTIONS on Information and Systems*, 91(4):959–968.

Sanchez-Graillet, O. and Poesio, M. (2004). Acquiring bayesian networks from text. In *LREC*.

Satpal, S., Bhadra, S., Sellamanickam, S., Rastogi, R., and Sen, P. (2011). Web information extraction using markov logic networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1406–1414. ACM.

Schulman, J., Levine, S., Moritz, P., Jordan, M. I., and Abbeel, P. (2015a). Trust region policy optimization. *CoRR*, abs/1502.05477.

Schulman, J., Moritz, P., Levine, S., Jordan, M. I., and Abbeel, P. (2015b). High-dimensional continuous control using generalized advantage estimation. *CoRR*, abs/1506.02438.

Sgouritsa, E., Janzing, D., Hennig, P., and Schölkopf, B. (2015). Inference of cause and effect with unsupervised inverse regression. In *Artificial Intelligence and Statistics*, pages 847–855.

Singh, S. P., Jaakkola, T. S., Jordan, M. I., et al. (1994). Learning without state-estimation in partially observable markovian decision processes. In *ICML*, pages 284–292.

Sloman, S. A. and Hagmayer, Y. (2006). The causal psycho-logic of choice. *Trends in Cognitive Sciences*, 10(9):407–412.

Spirtes, P. and Glymour, C. (1991). An algorithm for fast recovery of sparse causal graphs. *Social science computer review*, 9(1):62–72.

Spirtes, P., Glymour, C. N., and Scheines, R. (2000). *Causation, prediction, and search*. MIT press.

Spirtes, P., Meek, C., Richardson, T., and Meek, C. (1999). An algorithm for causal inference in the presence of latent variables and selection bias.

Stegle, O., Janzing, D., Zhang, K., Mooij, J. M., and Schölkopf, B. (2010). Probabilistic latent variable models for distinguishing between cause and effect. In *Advances in Neural Information Processing Systems*, pages 1687–1695.

Su, C., Andrew, A., Karagas, M. R., and Borsuk, M. E. (2013). Using bayesian networks to discover relations between genes, environment, and disease. *BioData mining*, 6(1):6.

Suppes, P. (1970). *A probabilistic theory of causality*. North-Holland Publishing Company Amsterdam.

Sutton, R. S. (1988). Learning to predict by the methods of temporal differences. *Machine learning*, 3(1):9–44.

Sutton, R. S. and Barto, A. G. (1998). *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge.

Teh, Y. W. (2011). Dirichlet process. In *Encyclopedia of machine learning*, pages 280–287. Springer.

Tsamardinos, I., Brown, L. E., and Aliferis, C. F. (2006). The max-min hill-climbing bayesian network structure learning algorithm. *Machine learning*, 65(1):31–78.

Waldmann, M. R., Cheng, P. W., Hagmayer, Y., and Blaisdell, A. P. (2008). Causal learning in rats and humans: A minimal rational model. *The probabilistic mind. Prospects for Bayesian cognitive science*, pages 453–484.

Wellen, S. and Danks, D. (2012). Learning causal structure through local prediction-error learning. Cognitive Science Society.

Xu, Z., Tresp, V., Yu, K., and Kriegel, H.-P. (2006). Learning infinite hidden relational models. *Uncertainity in Artificial Intelligence (UAI2006)*.

Xu, Z., Tresp, V., Yu, K., Yu, S., and Kriegel, H.-P. (2005). Dirichlet enhanced relational learning. In *Proceedings of the 22nd international conference on Machine learning*, pages 1004–1011. ACM.

Yerushalmy, J. and Palmer, C. E. (1959). On the methodology of investigations of etiologic factors in chronic diseases. *Journal of chronic diseases*, 10(1):27–40.

Yu, K., Tresp, V., and Yu, S. (2004). A nonparametric hierarchical bayesian framework for information filtering. pages 353–360.

Yu, X. and Lam, W. (2010). Jointly identifying entities and extracting relations in encyclopedia text via a graphical model approach. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 1399–1407. Association for Computational Linguistics.