Escalabilidad del descubrimiento de redes bayesianas causales mediante modelos de lenguaje a gran escala: metodología y análisis comparativo

Nikolay Babakov[1], Ehud Reiter[2], and Alberto Bugarín[1]

[1] Centro Singular de Investigacion en Tecnoloxías Intelixentes (CiTIUS)
Universidade de Santiago de Compostela
{nikolay.babakov,alberto.bugarin.diz}@usc.es
[2] University of Aberdeen
e.reiter@abdn.ac.uk

Resumen En este trabajo presentamos un resumen del trabajo "Scalability of Bayesian Network Structure Elicitation with Large Language Models: a Novel Methodology and Comparative Analysis" [2], presentado en la 31st International Conference on Computational Linguistics (COLING) y publicado en enero de 2025, y enriqueciéndolo con contenido adicional del trabajo "CausalGraphBench: a Benchmark for Evaluating Language Models Capabilities of Causal Graph Discovery" [1], para hacerlo más actualizado y autocontenido.

Keywords: Descubrimiento de Grafos Causales · Redes Bayesianas · Modelos de lenguaje a gran escala · Benchmarking

1. Introducción

En este trabajo presentamos un resumen de nuestro paper "Scalability of Bayesian network structure elicitation with large language models: a novel methodology and comparative analysis" [2], presentado en la 31st International Conference on Computational Linguistics (COLING), enriqueciéndolo con contenido adicional del trabajo [1], para hacerlo más actualizado y autocontenido, puesto que esta publicación, más reciente, preserva el método original descrito en [2], propone un nuevo método e incluye comparaciones más equitativas utilizando un banco de pruebas más exhaustivo.

Los grafos causales son fundamentales en modelos de razonamiento, como las Redes Bayesianas (RB), en los que las variables son los nodos y las dependencias causales directas se representan mediante arcos dirigidos. La estructura de las RB suele aprenderse a partir de los datos u obtenerse de expertos, pero recientemente se están usando modelos de lenguaje a gran escala (LLM) para inferir relaciones causales plausibles a partir de los nombres de las variables y del contexto del dominio. Esto solo resulta útil si los métodos pueden evaluarse de forma imparcial, algo que resulta difícil en la bibliografía, puesto que se emplean distintos conjuntos de grafos, indicaciones, protocolos de evaluación y supuestos de contaminación.

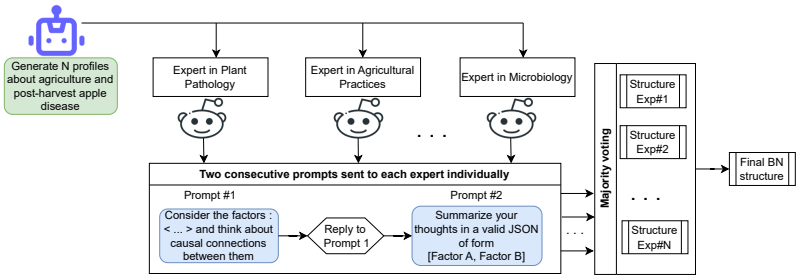


Figura 1. Resumen del método propuesto para el descubrimiento de RB causales.

Nuestra propuesta [2] describe un método de tipo Delphi basado en LLM para el descubrimiento de RB causales ("structural learning"), inspirado en la discusión colectiva entre expertos y la votación por mayoría. Adicionalmente, [1] presenta un banco de pruebas para evaluar este y otros métodos de descubrimiento de grafos causales ("Causal Graph Discovery", CGD) basados en LLM.

2. Métodos

Existen numerosos métodos que aprovechan las capacidades de los LLM para la tarea de CGD: i) el que usaremos como referencia ("baseline") utiliza un "prompt" con el contexto completo para generar todos los arcos, ii) el método Harness, adaptado de Ban et al. [3] pide al modelo que comprenda los nodos, infiera el grafo y revise los arcos propuestos, iii) el método de consultas por pares y tríos pregunta por cada par o trío de nodos [6]; estos métodos son costosos y su uso se limita a grafos pequeños y iv) el método eficiente sigue una estrategia de construcción iterativa con crecimiento lineal de las consultas [4].

Nuestro método adapta el protocolo Delphi al descubrimiento de RB causales mediante un LLM. Un LLM facilitador genera un número impar de perfiles de expertos relevantes para el ámbito en cuestión.Cada perfil recibe la lista de nodos y el contexto del grafo, y obtiene las relaciones causales directas plausibles devolviendo una lista de arcos. La estructura final incluye un arco solo cuando la mayoría de los expertos lo propone. Si un experto crea un ciclo o direcciones mutuamente inconsistentes, se resuelve antes de la agregación para que el grafo resultante siga siendo acíclico. La Figura 1 muestra la descripción general del método original. Su principal beneficio es su robustez frente a alucinaciones idiosincrásicas, mientras que su coste es el de realizar llamadas adicionales al LLM.

3. Banco de pruebas

Nuestro trabajo sobre el banco de pruebas CausalGraphBench formaliza el CGD de la siguiente manera: dado un conjunto de variables V y metadatos que aclaran su significado, el modelo debe generar un conjunto de arcos E^* que se

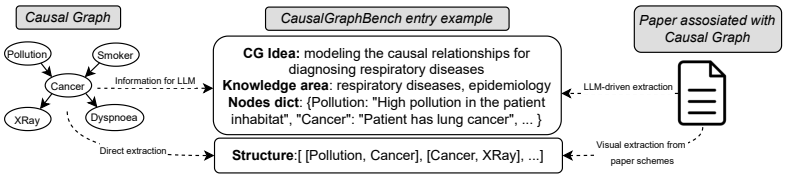


Figura 2. Ejemplo de creación de una entrada de referencia en CausalGraphBench.

Método	$\leq$10 nodos	11-50 nodos	Todos
Pair	1.67	–	–
Triplet	2.02	–	–
Efficient	1.16	1.66	–
Baseline	0.65	0.96	1.00
Harness	0.66	0.93	0.96
Delphi/proposed	0.80	1.07	1.06
Fine-tune	0.64	1.02	1.10

Cuadro 1. Resultados seleccionados de CausalGraphBench para GPT-4o, expresados como la SHD media normalizada (métrica decreciente). Los métodos de pares, tríos y los métodos eficientes se limitan a grafos de tamaño pequeño o mediano, debido a su coste.

aproxime al grafo objetivo $G = (V, E)$. El banco de pruebas contiene 35 CGD recopilados de repositorios públicos de RB y de trabajos académicos: 14 grafos de repositorios públicos, que suelen ser de mayor tamaño, y 21 grafos extraídos de artículos, que suelen ser más pequeños. Cada entrada incluye la descripción del grafo, el ámbito en el que se ha definido, un diccionario de descripción de los nodos y la lista de arcos objetivo. Los metadatos compartidos hacen que las comparaciones sean más equitativas, ya que cada método recibe la misma información en lugar de una indicación específica del método o un subconjunto del contexto. La Figura 2 muestra la figura original de construcción de entradas de CausalGraphBench.

4. Experimentos y conclusiones

En la experimentación hemos usado GPT-4o, Llama-3.3-70B-Instruct y Llama-3.1-8B-Instruct. La calidad de los grafos se mide mediante la distancia de Hamming estructural (SHD) [5] media entre ambos, normalizada por el número de arcos, lo que permite comparar grafos de diferentes tamaños.

La Tabla 1 presenta los resultados de los experimentos. El banco de pruebas modifica la interpretación del método original. En [2], comparamos Delphi con Harness y se mostró superior para uno de los LLM bajo estudio, aunque el rendimiento de ambas aproximaciones se degrada con el tamaño del grafo. En [1] se amplía la comparación: los métodos con el contexto completo a menudo mejoran el rendimiento de los de pares exhaustivos y de tripletas, que sobredimensionan

las relaciones locales y pierden el contexto global del grafo. Con este nuevo protocolo, también se muestra que Delphi no es superior de forma consistente a modelos más simples, como los de "prompting" de grafo completo o Harness. Esto, en todo caso, no invalida el método, sino que permite identificar cuándo sus llamadas adicionales resultan útiles y cuándo son redundantes. El rendimiento decrece con los métodos y los LLM en grafos de mayor tamaño, lo que confirma que la escalabilidad es el principal reto abierto en este ámbito.

La contribución combinada es tanto metodológica como evaluativa: proponemos un nuevo método de obtención basado en LLM y CausalGraphBench aclara su valor práctico al probarlo en un entorno de evaluación imparcial. Los resultados sugieren que conservar el contexto completo del grafo suele ser más eficaz que descomponer la tarea en múltiples llamadas locales, y que una mayor complejidad de las llamadas no resulta automáticamente beneficiosa. En particular, el método Delphi propuesto, entre otros, parece redundante en algunas llamadas, al compararlo con alternativas más simples que utilizan el contexto completo.

Agradecimientos

Este trabajo forma parte del proyecto de I+D+i TED2021-130295B-C33, financiado por el MCIN/AEI/10.13039/501100011033/ y por el programa "NextGenerationEU/PRTR de la Unión Europe." Los autores agradecen también el apoyo de la Consellería de Educación, Universidades y Formación Profesional de Galicia y del «FEDER Una forma de hacer Europa» a través de las ayudas "Acreditación del Centro de Investigación de Galicia 2024-2027 ED431G-2023/04."

Referencias

1. Babakov, N., Reiter, E., Bugarín, A.: CausalGraphBench: a benchmark for evaluating language models capabilities of causal graph discovery. In: Zhao, J., Wang, M., Liu, Z. (eds.) Proc. 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop) (Jul 2025)
2. Babakov, N., Reiter, E., Bugarín, A.: Scalability of Bayesian network structure elicitation with large language models: a novel methodology and comparative analysis. In: Rambow, O., Wanner, L., Apidianaki, M., Al-Khalifa, H., Eugenio, B.D., Schockaert, S. (eds.) Proc. 31st Int. Conf. on Computational Linguistics (Jan 2025)
3. Ban, T., Chen, L., Wang, X., Chen, H.: From query tools to causal architects: Harnessing large language models for advanced causal discovery from data. arXiv preprint arXiv:2306.16902v1 (2023). https://doi.org/10.48550/arXiv.2306.16902
4. Jiralerspong, T., Chen, X., More, Y., Shah, V., Bengio, Y.: Efficient causal graph discovery using large language models. arXiv preprint arXiv:2402.01207v4 (2024). https://doi.org/10.48550/arXiv.2402.01207
5. Tsamardinos, I., Brown, L.E., Aliferis, C.F.: The max-min hill-climbing bayesian network structure learning algorithm. Machine Learning **65**, 31–78 (2006)
6. Vashishtha, A., Reddy, A.G., Kumar, A., Bachu, S., Balasubramanian, V.N., Sharma, A.: Causal inference using LLM-guided discovery. arXiv preprint arXiv:2310.15117v1 (2023). https://doi.org/10.48550/arXiv.2310.15117