Insurance
Redes Bayesianas

Dany Efrain Rubiano Jiménez
Universidad de Santiago de Chile

Abstract. Insurance es un estudio de las características a tener en consideración en el contexto de los seguros para automóviles, tomando en cuenta las primas asociadas a robos, accidentes, costos médicos entre otros. Haciendo uso del método de Redes Bayesianas, se busca modelar este problema y obtener conocimiento a partir de consultas con métodos de propagación de la evidencia.

Keywords: Redes Bayesianas, seguros, grafo dirigido.

1 Introducción

Las redes bayesianas unen los conocimientos específicos de la estadística, la teoría de grafos y la optimización. Componen un método capaz de analizar problemas en donde existe el conocimiento de que no todos los atributos están en la misma causalidad, a manera de identificar el aporte o incidencia de ciertas características sobre otra característica o clase.

Durante esta experiencia, se procede a la aplicación de las redes bayesianas sobre el conjunto de datos de características a considerar en los seguros de autos, con el fin de inferir la relevancia de cada una de las variables y obtener conocimiento a partir de la realización de diferentes consultas en función de las características consideradas para las diferentes primas que se pueden desprender.

2 Métodos y datos

2.1 Métodos utilizados

El método de redes bayesianas hace uso de la base probabilística. La información cuantitativa de la red viene dada por la probabilidad a priori de los nodos que no tienen padres P(padres) y de la probabilidad condicional de los nodos con padres P(hijo|padres). El método permite calcular, a partir de los datos anteriores, las probabilidades a posteriori de una evidencia observada utilizando el Teorema de Bayes. El problema surge, al igual que en un clasificador bayesiano ingenuo, cuando se requiere obtener la probabilidad condicional de P(hijos/padres) cuando los padres son muchos significando un cálculo complejo, sin embargo, el método hace uso del supuesto de independencia condicional, simplificando el cálculo probabilístico. [1]

En primer lugar, se procede a la aplicación de la red bayesiana asociada al problema, variando el método de la construcción. Se usan para ello los métodos de

Mediante el criterio del BIC, se discriminan los métodos seleccionando el mejor. Luego se modela la red bayesiana en base a diferentes configuraciones de los parámetros de la función y con la variación de la disposición de los datos, tratando de encontrar un modelo semejante al que se encuentra disponible en la literatura.
En base a los datos, se establecen los pesos de las conexiones entre el grafo dirigido que representa la red, a fin de realizar en una última instancia diferentes consultas según el modelo obtenido en comparación con el modelo original.

2.2 Datos utilizados

El conjunto de datos contiene diferentes variables asociadas a características a tener en cuenta en los seguros de autos. Las variables totales corresponden a 27 y sus respectivos nombres se refleja en la tabla 2.1. El conjunto de datos cuenta con 2000 ejemplos, para los que no se registran datos faltantes o nulos.

Tabla 2.1: *Variables del conjunto de datos.*

Age	ExtraCar	Cushioning
SocioEcon	VehicleYear	MedicalCost
GoodStudent	Antilock	LiabilityCost
RiskAversion	Ruggedeness	OtherCarCost
SeniorTrain	AntiTheft	PropertyCost
DrivingSkill	HomeBase	OwnCarCost
Mileage	CarValue	OwnDanage
MakeModel	AirBag	DrivQuality
DrivingHist	Accident	Airbag

3 Resultados

En primer lugar, se utilizan diferentes métodos para la construcción de los grafos correspondientes a las redes bayesianas. Los resultados en términos de los verdaderos positivos (tp, arcos idénticos en el modelo original y el generado), falsos positivos (fp, arcos no presentes en el modelo original), falsos negativos (fn, arcos con cambios de sentido con respecto al original) y del criterio BIC se presentan en la tabla a continuación.

Tabla 3.1: *Contraste de los resultados de los métodos para generar la red bayesiana.*

	tp	**fp**	**fn**	**BIC**

HC	26	26	24	-266113.0284
mmhc	15	37	6	-300926.9492
mmpc	0	52	21	error

Obteniendo la mejor configuración en base a la variación de la semilla y a la variación del parámetro *restart*, se procede a realizar distintas pruebas variando la *whitelist* y la *blacklist*, intentando mejorar el modelo en términos del criterio BIC y de los verdaderos positivos. La tabla 3.2 refleja los resultados de la mejor red bayesiana encontrada. Cabe destacar que el BIC del modelo original corresponde a -265858,559.

Tabla 3.2: *Resultados de la mejor red bayesiana encontrada.*

	tp	**fp**	**fn**	**BIC**
HC	52	0	0	-264819.9370

A modo comparativo, se presenta en la figura 3.1 el modelo generado (en la posición izquierda) en contraste con el modelo original, en donde las líneas de color rojo representan la similitud entre los arcos.

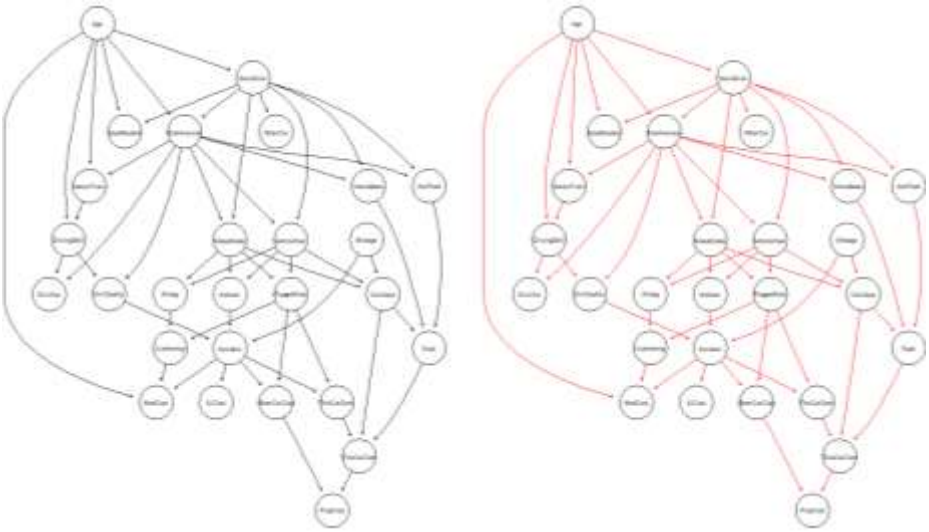

Figura 3.1*: Contraste de las redes bayesianas (en la posición izquierda se encuentra el modelo generado, mientras que a la derecha está el modelo original).*

Para la posterior realización de consultas, se registran los pesos respectivos de las conexiones de los grafos dirigidos correspondientes al modelo generado (posición izquierda) y al modelo original.

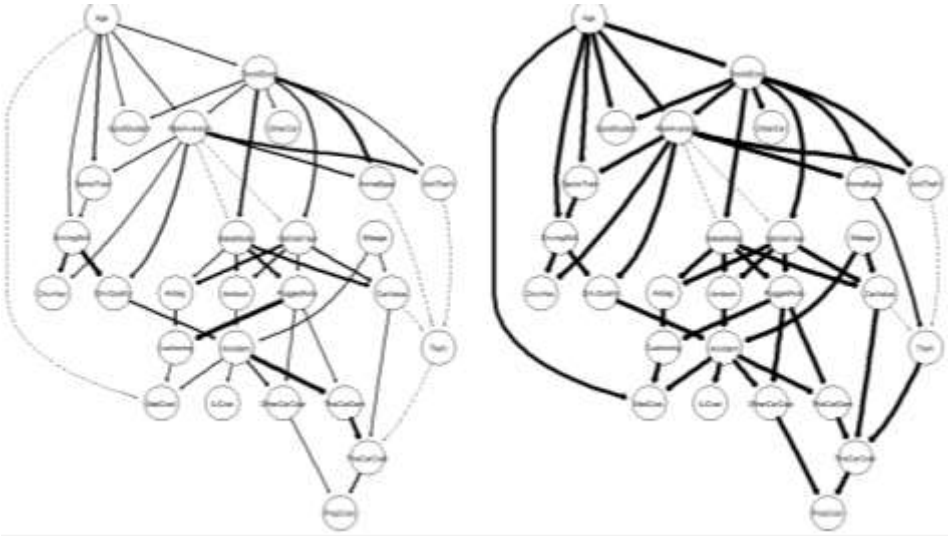

***Figura 3.2**: Contraste de los pesos de los arcos de las redes bayesianas (en la posición izquierda se encuentra el modelo generado, mientras que a la derecha está el modelo original).*

A partir del modelo generado, se procede a realizar diferentes consultas con el fin de obtener conocimiento del problema en cuestión. En primer lugar, se aplican consultas generales, tomando en cuenta los nodos principales identificados en el modelo. En específico se toma como nodo principal de evidencia a la edad, mientras que para el evento de la consulta se toman los nodos correspondientes a los costos médicos, accidentes, aversión de riesgo, y costos proporcionales.

Las consultas generales propuestas con base todas en la edad son:

1. ¿De qué manera incide la edad en tener costos médicos altos?
2. ¿En qué medida afecta la edad en generar costos proporcionales millonarios?
3. ¿Cómo incide la edad en la ocasión a tener un accidente severo?
4. Dada la aversión aventurera, ¿cómo esta se asocia a la edad?

***Tabla 3.3:** Probabilidades de las consultas generales en base a la edad.*

Consulta	Modelo Generado			Modelo Original		
	Adolescent	**Adult**	**Senior**	**Adolescent**	**Adult**	**Senior**
1	2,907%	2,213%	2,263%	3,341%	1,895%	1,213%
2	2,550%	1,609%	1,103%	2,564%	1,486%	0,976%
3	17,35%	10,94%	6,637%	17,466%	10,91%	7,440%
4	49,85%	25,06%	9,430%	47,39%	24,81%	8,080%

Otras consultas más específicas en base al modelo, considerando varios parámetros y los pesos de las conexiones entre los nodos, son presentadas a continuación:

1. *¿En qué medida intervienen en un accidente severo características de kilometraje, antilock y calidad de conducción?*
   Se encuentra que en los accidentes severos, según el modelo generado, la característica que más incide entre las consideradas es la de la calidad pobre de conducción, esto a pesar de la variación del kilometraje. Por su parte la presencia del sistema de frenos antilock, disminuye en un gran porcentaje la probabilidad de accidente severo. El valor más alto de probabilidad corresponde a la no presencia de antilock, calidad de conducción pobre y un kilometraje domino, con un 45,57%. Por su parte el modelo original para este mismo caso presenta una probabilidad de 49,89% y las inferencias de la consulta son similares.

2. *Dado un accidente, ¿en qué medida la presencia o no de airbag y las condiciones de amortiguación generan costos médicos altos?*
   Aquí se observa que los costos médicos altos dado un accidente severo o moderado son mayormente condicionados por un sistema de amortiguación pobre o justo, no siendo relevante la presencia de airbag. La probabilidad más alta se da en el caso de un accidente severo, amortiguación pobre y presencia de airbag con un 26,85%. El modelo original para este caso presenta una probabilidad de un 22,11%.

3. *Dada la edad y la situación socioeconómica, ¿cómo estas inciden en la aversión psicópata?*
   En esta oportunidad se da el caso que en el modelo generado, los adolescentes con una situación económica de riqueza son lo que tienen mayor probabilidad de aversión psicópata con un 3,03%, y esto se generaliza a los adolescentes de cualquier situación socioeconómica. Por su parte, en el modelo original de la literatura son los adultos de una situación económica de riqueza quienes tienen mayor probabilidad de presentar una aversión psicópata con un 2,88%.

4. *Ante un robo de un auto, ¿en qué medida afecta el valor del auto, la presencia de un sistema antirrobos y la ubicación del hogar?*
   En este caso son muy pocos los casos donde hay robos registrados, siendo solo 31 de los 2000 registrados. Teniendo esto en cuenta, se presenta que la probabilidad más alta se da en el caso de la ubicación del hogar en la ciudad, la presencia de un sistema de antirrobos y con un valor del auto de 20000 dólares, con un 2,09%. Para el modelo original se da el mismo caso con una probabilidad de 1,02%.

4 Discusión

En base a los distintos resultados expuestos se puede afirmar que no existe una diferencia significativa entre el modelo generado y el modelo original ofrecido por la literatura. Esto se demuestra en que la cantidad de verdaderos positivos del modelo generado es igual al número de arcos presentes en ambos modelos. A partir de ello se puede inferir que el problema es de cierta manera determinista ya que con la red bayesiana se alcanza el mismo modelo que el de la literatura.
En la perspectiva del criterio BIC se puede observar que el modelo generado tiene una cierta mejora, pero no muy significativa.
Las principales diferencias entre ambos modelos se denotan en torno a los pesos de las conexiones de los grafos dirigidos correspondientes a cada red, sin embargo, dados los resultados de las diferentes consultas realizadas, no se encuentran mayores diferencias. Sólo se presentan inferencias diferentes a la consulta de la influencia de la edad y la situación socioeconómica en la aversión psicópata, en donde para el modelo generado se dan mayores probabilidades en los adolescentes, mientras que en el modelo original se da en los adultos.
En lo que respecta a la obtención de conocimiento a partir de la red, el nodo principal corresponde a la edad, que dado el tipo de problema, es el primer elemento a tomar en cuenta para otorgar un seguro, fijar la prima correspondiente, o toda aquella acción a tener en cuenta para una aseguradora. Esto en función de los nodos de evento correspondientes a costos médicos, costos proporcionales, accidentes, entre otros nodos intermedios, que pueden determinar las primas por cada tipo de seguro u otra información de interés.

5 Conclusiones

Las redes bayesianas son una poderosa herramienta para la extracción de conocimiento, permitiendo representar significativamente un problema mediante un grafo dirigido con la correspondiente fuerza entre las conexiones y realizar las consultas pertinentes. Al comparar este método con otros usados para la minería de datos como las redes neuronales, las redes bayesianas ofrecen la ventaja de ser entendibles, capaces de modelar por el usuario y de soportar inferencias de cualquier tipo de pregunta, mientras que las redes neuronales funcionan más bien como una caja negra. La principal dificultad de las redes bayesianas radica en la complejidad del método en sí.

La construcción de las redes bayesianas debe ser vigilada, teniendo en cuenta para su modelamiento un juicio de expertos o un modelo ya comprobado, de manera que la extracción de conocimiento pueda ser validada.
A través de las configuraciones correctas de los parámetros para la construcción del modelo y de el orden en que los datos son asociados, dependiendo del tipo de problema, es posible obtener los mismos resultados que un modelo construido por juicio de expertos.

6 Referencias

1. Chacón, M. Taller de minería de datos avanzada, Capitulo IV: Redes Bayesianas
2. Binder, J. (1997) Adaptive Probabilistic Networks with Hidden Variables