Redes Bayesianas aplicadas para dataset de meteorología *hailfinder*

Pedro Pablo Silva Antilef[1] [0000−1111−2222−3333]

Universidad de Santiago de Chile, Facultad de ingeniería, Departamento de ingeniería informática, Estación Central, Santiago de Chile pedro.silva@usach.cl

Abstract. En el presente trabajo se presenta un análisis del *dataset* denominado *Hailfinder*, el cual es un conjunto de datos tomados por instrumento y por apreciación de expertos relacionadas al tiempo en el año 1989 para el norte de California. Estos datos fueron obtenidos por el Laboratorio de predicción de sistemas del *National Oceanic and Atmospheric Administration* (NOAA/FSL). Este dataset fue utilizado para predecir la presencia tiempo extremo para la zona de estudio utilizando Redes Bayesianas. En este trabajo se analizan las relaciones generadas sobre este conjunto (redes bayesianas) de datos utilizando los algoritmos de optimización *Hill-Climbing*, *Max-Min Hill-Climbing* y *Max-Min Parents and Children*. También se analizaron las probabilidades condicionales de las relaciones, pudiendo inferir que la variable que influye de manera más directa en la aparición de cualquier tiempo extremo en la región es la oscuridad de las nubes.

Keywords: Bayessian network · Optimization algorithms · Data Minning.

1 Introducción y estado del arte

La predicción es una de las partes que entregan más valor dentro los crecientes campos de *Data Science*, tanto en el sector público, privado y académico. Es en este contexto que es necesario poder conocer y entender las relaciones que existen en los datos, con el fin de tomar decisiones en el modelamiento de manera "sabia" y obtener valores cercanos a la realidad. En este sentido las redes bayesianas son una buena manera de obtener información de los dataset y las relaciones de sus datos, al tratarse de un modelo probabilístico gráfico, el cual nos permite ver de manera visual y en términos de probabilidades condicionales de las relaciones entre variables [1].

Los modelos probabilísticos gráficos han sido ampliamente utilizados en la literatura en distintas áreas de aplicación. Hablando específicamente de las redes bayesianas, podemos encontrar un ejemplo de aplicación en ecología [2], utilizada para sintetizar, predecir y analizar la incertidumbre del modelo tomando en cuenta data de varios procesos relacionados a la eutrofización (proceso en el cual un cuerpo de agua se convierte progresivamente en un ambiente rico en minerales

y nutrientes, aumentando la probabilidad de proliferamiento de fitoplanton) en el estuario del Río Neuse.

Otra aplicación de las redes bayesianas, es la predicción meteorológica, una de esas aplicaciones es la realizada por [3], quienes predicen las lluvias en el norte de España a partir de información espacio-temporal obtenida de estaciones de monitoreo de condiciones meteorológicas sumado al conocimiento de expertos. Finalmente lograron demostrar que al tomar en consideración la variable geográfica lograron mejorar los modelos clásicos que no contemplaban esta dimensión. Ota aplicación a la predicción del tiempo es precisamente la llevada a cabo por [4], quien en su trabajo busca predecir la presencia de granizos en el norte de California a partir de información proporcionada por el Laboratorio de predicción de sistemas del *National Oceanic and Atmospheric Administration* (NOAA/FSL) aplicando redes bayesianas. Esta aplicación fue denominada como *Hailfinder*. Este último autor publicó la base de datos con la que se construyeron los modelos y es precisamente este *dataset* con el cual se trabajará en el presente trabajo. Para esto se utilizará el*software* estadístico R.

2 Datos

El *dataset* utilizado es una base de datos pública llamada *Hailfinder* o "buscador de granizos", el cual consta de 20.000 registros con un total de 56 columnas, todas relacionadas a variables ambientales obtenidas por un muestreo diario de manera y también apreciaciones subjetivas aportadas por expertos. De esas 56 variables, existen algunas relacionadas al output de los modelos realizados, es decir predicciones meteorológicas divididas en 3 categorías:

- Severo (SVR): Ocurrencia en algún punto en un área específica, en un periodo de tiempo específico de (i) granizo con un diámetro mayor o igual a 0.75 pulgadas, (ii) vientos superficiales de 50 nudos o más, o (iii) un tornado.
- Significante (SIG): Ocurrencia en algún punto en un área específica, durante un periodo de tiempo específico, de al menos (i) granizos de un diametro entre 0.25 y 0.74 pulgadas, (ii) vientos superficiales de entre 35 y 49 nudos, (iii) lluvias de al menos 2 $pulgadas \cdot h^{-}1$, o (iv) un embudo de nubes.
- Nil (XNIL): Ausencia de significancia o tiempo extremo.

Un dato importante a considerar que el área de estudio se dividió en 4 regiones, donde la región 1 es la región montañosa, mientras que las otras regiones son consideradas planas, además se contempla una región 5, el cual corresponde a la unión de las 4 primeras.

Otro concepto importante de mencionar son las relacionadas a los escenarios, estos se derivan desde la clasificación realizada por John Brown, uno de los autores del articulo, quien es un experto en meteorología de Colorado. Estos escenarios buscan describir aproximadamente un 80% de los días típicos, obviamente sujetos a la experiencia del la persona [4].

3 Metodología

Como ya se explicó anteiormente, esta muestra de datos se analizará con redes bayesianas. Una de las métricas utilizadas es BIC o criterio de información bayesiano, el cual es utilizado como criterio para la selección de modelos. Una de las grandes ventajas de este método es que se puede interpretar de manera gráfica, es decir las interacciones se muestran de manera explícita en un grafo acíclico dirigido. En este contexto se busca obtener información de la red bayesiana calculada con diferentes algoritmos de búsqueda como los son *Hill-Climbing* (hc), *Max-Min Hill-Climbing* (mmhc) y *Max-Min Parents and Children* (mmpc). La idea es analizar el rendimiento de estos algoritmos en las redes bayesianas para luego proponer algunas modificaciones y obtener un mejor resultado del modelo.

Se definieron dos experimentos, que variarán principalmente en los tipos de algoritmos de optimización para el *dataset* completo y para el *dataset* excluyendo algunas variables, principalmente aquellas variables que luego se combinaban en una nueva variable, como el caso de la característica "movimiento vertical", la cual tiene 3 variables que finalmente se integran en la columna CombVerMo (movimiento vertical combinado). En total se eliminan 12 de las 56 variables: *IRCloudCover, VISCloudCov, N07muVerMo, SubjVertMo, QGVertMotion, SatContMoist, RaoContMoist, VISCloudCov, IRCloudCover, LowLLapse, MeanRH, MidLLapse*

4 Resultados

En esta sección se presentan los resultados obtenidos de los diferentes experimentos planteados y explicados en la sección 3. Estos resultados serán discutidos en la siguiente sección.

A continuación se muestran los resultados obtenidos para cada uno de los experimentos:

Experimento	**Algoritmo**	**BIC**
1	hc	-990474,8
	mmhc	-1144947
	mmpc	-
2	hc	-785273.9
	mmhc	-917698.7
	mmpc	-

Table 1: Tabla con resultados obtenidos para cada uno de los experimentos

A partir de los dos experimentos, se termina seleccionando los modelos con mejores resultados según la métrica BIC, es decir el experimento 1 y 2 utilizando el algoritmo *Hill-Climbing*.

Desde acá es posible obtener aun más conocimiento de las redes bayesianas creadas, principalmente a través de la obtención de tablas de probabilidad condicionales mediante el algoritmo de Máxima expectación o propagación de la evidencia. Para el caso del presente informe, interesa analizar la tabla de probabilidades para las variables de predicción, con el fin de saber cómo se ven afectadas por sus relaciones en el grafo. Para esto se obtiene la tabla para las variables *R5Fcst*, donde podemos obtener las siguientes tablas:

	MountainFcst		
R5Fcst	**SIG**	**SVR**	**XNIL**
SIG	1	0	1
SVR	0	1	0
XNIL	0	0	0

Table 2: Probabilidad condicional para R5Fcst con N34StarFcst = SIG

	MountainFcst		
R5Fcst	**SIG**	**SVR**	**XNIL**
SIG	0	0	0
SVR	1	1	1
XNIL	0	0	0

Table 3: Probabilidad condicional para R5Fcst con N34StarFcst = SVR

	MountainFcst		
R5Fcst	**SIG**	**SVR**	**XNIL**
SIG	1	0	0
SVR	0	1	0
XNIL	0	0	1

Table 4: Probabilidad condicional para R5Fcst con N34StarFcst = XNIL

Finalmente, se pueden hacer consultas a la red bayeseiana ya ajustada y con la propagación de la evidencia realizada. Para esto se debe seleccionar un evento y una evidencia, con el fin de obtener una probabilidad condicional que nos permita extraer más información de esta red.

Las consultas realizadas fueron las siguientes:

Evento	**Evidencia**	**Probabilidad**
R5Fcst = SVR	CombMoisture = VeryWet	0.3056243
	CombVerMo = StrongUp	0.329092
	CombClouds = Cloudy	0.3031066
	LLIW = Strong	0.3389956
	CldShadeConv = Marked	0.3922674
	WndHodograph = DCVZFavor	0.3183664

Table 5: Probabilidades condicionales consultadas a la red

5 Discusión y conclusiones

Las redes bayesianas son una herramienta potente para el análisis de datos con una gran cantidad de variables, como el caso del *dataset Hailfinder*, donde se tienen muchas variables ambientales que para alguien no experto pueden ser difíciles de entender y por ende de extraer información valiosa. La utilización de un grado acíclico dirigido es una excelente manera de entender las relaciones entre esa gran cantidad de variables. Por otra parte la métrica de BIC fue útil para poder elegir los mejores modelos generados a partir de la combinación de variables.

Entre los tres algoritmos de optimización *Hill-Climbing* (hc), *Max-Min Hill-Climbing* (mmhc) y *Max-Min Parents and Children* (mmpc), se obtuvieron relaciones nulas entre los nodos para el último algoritmo y ́las mejores métricas para el primer y más simple algoritmo de búsqueda.

El procesamiento constó simplemente de la generación de una "lista blanca" o selección de variables adecuadas, un procesamiento recomendado en el artículo principal que sustenta esta red, ya que elimina data que podría ser redundante, sobre estimando ciertas dimensionalidades del fenómeno estudiado, en este caso la presencia de clima extremo. Este procesamiento generó un aumento significativo entre el experimento 1 y el experimento 2, tal como se puede apreciar en la Tabla 1.

Por otra parte, el análisis de probabilidad condicional para el nodo *R5Fcst* (Predicción para la región 5), nos otorga 3 tablas, donde la probabilidad de sus posibles valores se condicionan por los posibles valores del nodo *MountainFcst*. En este sentido, se puede apreciar que cuando existe presencia de tiempo extremo en la región plana, se considera que si o si existe tiempo extremo en la zona montañosa, como se puede apreciar en la Tabla 3. Por otro lado, si existe un tiempo significativo en la región plana, existe un 100% de probabilidad de que también sea significante en la región montañosa (Tabla 2).

Por último se generaron varias consultas a la red bayesiana del experimento 2, con la intención de analizar cual es la probabilidad condicional de que en algún lugar de la región exista clima extremo, teniendo como evidencia varios criterios, tal como se muestra en la Tabla 5. En este contexto, casi todas las combinaciones tuvieron probabilidades cercanas al 30%, como por ejemplo que la variable de combinación de humedad tuviera un valor "muy humedo" obtuvo un valor de 0.3, la variable de combinación de medidas de movimiento vertical con valor "Fuerte ascendiente" tuvo 0.329, variable de combinación de métricas de nubosidad con valor "nuboso" con una probabilidad del 30%, la variable de índice de tiempo extremo LLIW con valor "strong" obtuvo 0.338, la variable de hodografía de la nube (rosa de los vientos) con valor "DCVZFavor" obtuvo una probabilidad de 0.31. La única variable que parece ser más influyente en la aparición de tiempo extremo es la oscuridad de la nube cuando tiene el valor de "markado".

Si bien los resultados obtenidos tienen sentido según lo visto en el *paper* principal, no se logró realizar la división del *dataset* para analizar el rendimiento de la red, lo cual puede darnos pistas de lo bueno que pueden ser estos modelos

respectos a los experimentos realizados, pero no respecto a un panorama general (es decir comparación con otros modelos de la literatura). Es por esto que se recomienda realizar este procesamiento para analizar el rendimiento del modelo a través de varias métricas clásicas de aprendizaje de máquinas.

Si bien no se tiene la ubicación exacta ni aproximada de las mediciones que componen esta base de datos, si se hace una distinción entre áreas planas, áreas montañosas y todas las áreas. En este sentido podría ser interesante analizar este problema de manera separada para las observaciones en montaña y las observaciones en terreno plano, ya que probablemente se puedan extraer relaciones o características específicas de algún tipo de terreno.

References

1. Nagarajan, R., Scutari, M., & Lèbre, S. (2013). Bayesian networks in r. Springer, 122, 125-127.
2. Borsuk, M. E., Stow, C. A., & Reckhow, K. H. (2004). A Bayesian network of eutrophication models for synthesis, prediction, and uncertainty analysis. Ecological Modelling, 173(2-3), 219-239.
3. Cofino, A. S., Cano, R., Sordo, C., & Gutierrez, J. M. (2002). Bayesian networks for probabilistic weather prediction. In 15th Eureopean Conference on Artificial Intelligence (ECAI).
4. Abramson, B., Brown, J., Edwards, W., Murphy, A., & Winkler, R. L. (1996). Hailfinder: A Bayesian system for forecasting severe weather. International Journal of Forecasting, 12(1), 57-71.
5. Souza, T. L., Nishijima, M., & Fava, A. C. (2019). Do consumer and expert reviews affect the length of time a film is kept on screens in the USA?. Journal of Cultural Economics, 43(1), 145-171.
6. "https://blog.datumbox.com/machine-learning-tutorial-the-max-entropy-text-classifier/". Rescatado el 10 de Mayo de 2020.

6 Anexos

1. Red Bayesiana para *dataset* HailFinder completo (experimento 1), utilizando algoritmo *Hill-Climbing*

2. Red Bayesiana para *dataset* HailFinder con variables seleccionadas (experimento 2), utilizando algoritmo *Hill-Climbing*