Análisis probabilístico mediante redes bayesianas para el cálculo de la probabilidad de lluvia en diferentes localidades

Guillermo De la Torre-Gea y Oscar Delfín-Santiesteban

Tecnológica de Corregidora, Municipio de Corregidora, Santiago de Querétaro, México
gtorre.utcorregidora@gmail.com

Resumen La comprensión de las relaciones entre los factores climáticos contribuye a un mejor conocimiento de los fenómenos atmosféricos. Dichos factores se estudian mediante el diseño de modelos predictivos, sin embargo, las técnicas estadísticas convencionales no toman en cuenta las relaciones de dependencia entre estos factores. Los modelos de aprendizaje bayesianos consisten en la cuantificación de la probabilidad condicional, dando como resultado la identificación de relaciones causales entre las variables. En este trabajo se empleó un modelo de redes bayesianas para realizar el análisis probabilístico que nos permitió determinar las dependencias espaciales y temporales entre las precipitaciones de diferentes localidades que no son observables con otros métodos. Tomando en cuenta un conjunto de datos incompletos de la precipitación pluvial en 18 estaciones meteorológicas a partir de tres años, las relaciones de dependencia entre dichos valores de precipitación se pueden observar y calcular las probabilidades de lluvia condicionales.

Palabras clave: Modelos probabilísticos, clima, predicción, minería de datos, K2.

1. Introducción

La climatología se basa en un análisis estadístico de la información meteorológica medida y almacenada, las variaciones temporales que se producen en los parámetros climáticos se incorporan a los promedios estadísticos [14]. De esta forma, los factores que conforman el clima se estudian para desarrollar modelos de predicción. Para este propósito, es necesario obtener conjuntos de datos mediante un método sistémico y homogéneo, a partir de las estaciones meteorológicas durante períodos por lo menos de 30 años para ser considerados representativos. Las bases de datos climáticas contienen propiedades estadísticas sobre la precipitación pluvial para cada localidad, mientras que análisis conjunto de estas bases de datos podría contener las relaciones entre las precipitaciones de diferentes localidades. Así, esta información puede ser utilizada para analizar los problemas relacionados con la dinámica de la atmósfera y los posibles efectos que pueden causar en la climatología de las regiones. Sin embargo, las técnicas estadísticas en estos problemas, incluidos los métodos de regresión lineal para el pronóstico del tiempo, los métodos de agrupamiento y análisis de componentes principales para la identificación de los patrones atmosféricos de representación, contienen información fragmentada y asumen independencias espaciales ad-hoc

con el fin de obtener modelos simplificados [4], [2]. Por otro lado, los datos sobre precipitación pluvial obtenidos de las estaciones climáticas son a menudo incompletos, lo que obliga al uso de técnicas de minería de datos para el análisis [1], [11]. Por estas razones, es necesario aplicar técnicas de análisis que no afecten las propiedades inherentes de dichos datos.

Los modelos numéricos de predicción del clima son representaciones abstractas del mundo real. Estos modelos discretizan los datos utilizando funciones de aproximación para describir el comportamiento de las variables climáticas de interés en los estudios de predicción.

Hoy en día, los modelos numéricos son indispensables para la predicción climática. De acuerdo con [15], los parámetros en escala de tiempo variable podrían ser estimados utilizados para evaluar si existen tendencias estadísticamente significativas en datos climáticos a través de redes bayesianas (RB). Por otra parte, [21] propuso un modelo estocástico del espacio-tiempo que representa las dependencias temporales y espaciales de ocurrencia de precipitación por día. Un modelo RB para estimar la ocurrencia de precipitación basado en cadenas ocultas de Markov fue desarrollado por [10] utilizando el método de estimación de máxima probabilidad con datos incompletos. Según [13], es recomendable el empleo de RB sobre todo en conjuntos pequeños de datos disponibles.

El objetivo de este estudio es demostrar que las RB pueden ser empleadas para encontrar un modelo que mejor describa las relaciones entre la presencia de precipitación pluvial en distintas estaciones meteorológicas, mediante datos disponibles incompletos y limitados en el tiempo, con lo cual realizar el pronóstico de precipitación calculando las probabilidades condicionales de ocurrencia de precipitación pluvial.

2. Teoría de las RB

Las RB son representaciones del conocimiento desarrollados en el campo de la inteligencia artificial para el razonamiento aproximado [18], [16] y [5]. Una RB es un gráfico acíclico cuyos nodos corresponden a conceptos o variables, y cuyos enlaces definen las relaciones o funciones entre dichas variables [3]. Las variables se definen en un dominio discreto o cualitativo, y las relaciones funcionales describen las inferencias causales expresadas en términos de probabilidades condicionales, como se muestra en la Ecuación 1.

$$P(x_1,...,x_n) = \prod_{i=1}^{n} P(x_i \mid \text{padres}(x_i)) \qquad (1)$$

Las RB pueden ser usadas para identificar las relaciones entre las variables anteriormente indeterminadas o para describir y cuantificar estas relaciones, incluso con un conjunto de datos incompletos [11] y [19]. Los algoritmos de solución de las RB permiten el cálculo de la distribución de probabilidad esperada de las variables de salida. El resultado de este cálculo es dependiente de la distribución de probabilidad de las variables de entrada. A nivel mundial, las RB puede ser percibida como una

distribución de probabilidades conjunta de una colección de variables aleatorias discretas [7].

La probabilidad a priori $P(c_j)$ es la probabilidad de que una muestra x_i pertenezca a la clase C_J, definida sin ninguna información sobre sus valores característicos, como se muestra en la Ecuación 2.

$$P(c_j / x_i) = P(x_i / c_j)P(c_j) / \sum P(x_i / c_k)P(c_k) \quad (2)$$

Las máquinas de aprendizaje, en la inteligencia artificial, está estrechamente relacionado con la minería de datos, métodos de clasificación o agrupamiento en estadística, razonamiento inductivo y reconocimiento de patrones. Métodos estadísticos de aprendizaje automático se pueden aplicar al marco de la estadística bayesiana, sin embargo, la máquina de aprendizaje pueden emplear una variedad de técnicas de clasificación para producir otros modelos de RB [17] y [20]. El objetivo del aprendizaje mediante RB es encontrar el arreglo que mejor describa a los datos observados. El número de posibles estructuras de grafos acíclicos directos para la búsqueda es exponencial al número de variables en el dominio, el cual se encuentra definido por la Ecuación 3:

$$f(n) = \sum_{i=1}^{n} (-1)^{i+1} C^{n} 2_{i}^{(n-i)} f(n-i) \quad (3)$$

El algoritmo K2 constituye el método más representativo entre las aproximaciones de "búsqueda y resultado". El algoritmo comienza asignando a cada variable sin padres. A continuación, agrega de manera incremental los padres a la variable actual que aumenta su puntuación en la estructura resultante. Cuando cualquier adición de una madre soltera no puede aumentar la cuenta, deja de agregar padres a la variable. Desde un pedido de las variables conocido de antemano, el espacio de búsqueda bajo esta restricción es mucho menor que el espacio tomando la estructura entera, y no hay necesidad de comprobar los ciclos en el proceso de aprendizaje. Si el orden de las variables es desconocida, se puede realizar la busqueda en los ordenamientos [8].

3. Materiales y métodos

El área de estudio abarca el estado de Querétaro, México, que tiene una superficie de 11.689 km2 divididos en tres zonas climáticas diferentes: la parte sur, que cubre la provincia fisiográfica del Eje Neovolcánico, la región central que comprende las áreas de la región del Eje Neovolcánico, la Sierra Madre Oriental y la Mesa Central y la zona norte que corresponde a una porción de la Sierra Madre Oriental, como se muestra en la Fig. 1.

Se eligieron 18 estaciones climáticas en el estado de Querétaro para obtener los datos de precipitación pluvial, que representan los tres climas diferentes (Fig. 2 y 3), ubicando a Jalpan que corresponde a un clima de ACw, las estaciones meteorológicas que corresponde a un clima BS1k son: Cadereyta, Ezequiel Montes, Colón y Tequisquiapan, mientras que Amealco, San Joaquin y Huimilpan corresponden a un clima Cw de acuerdo a la clasificación de Koppen [6], como se muestra en la Fig. 2. Las

Fig. 1. Fisiografía del estado de Querétaro, fuente [12].

Fig. 2. Distribución climática del estado de Querétaro, fuente [12].

estaciones meteorológicas Corregidora, El Marques, CEA, El Milagro, Centro Sur, Centro Cívico, Rochera y Altamira corresponden a climas de transición.

Se obtuvieron datos de cada día durante el período comprendido entre el 01/01/2007 al 31/12/2010, a través de la red de estaciones meteorológicas de la Comisión Estatal de Aguas del estado de Querétaro, cuya ubicación se presenta en la Fig. 3, mediante el portal:

http://www.wunderground.com/weatherstation/ListStations.asp?selectedCountry=Mexico.

El conjunto de datos estaba compuesto por valores máximos, promedio y mínimos de precipitación pluvial, los cuales fueron discretizados en 2 valores: presencia y ausencia de lluvia.

Fig. 3. Estaciones meteorológicas en el estado de Querétaro, fuente [12].

3.1. Análisis bayesiano

El análisis fue realizado mediante el sistema ELVIRA versión 0.162 en tres etapas sugeridas por [16]:

1. Pre-proceso y tratamiento de datos: Se empleó el algoritmo de imputación por promedios, el cual sustituye los valores perdidos/desconocidos con la media de cada variable. No necesita parámetros. La discretización de las variables fue de tipo masiva, empleando el algoritmo Equal frecuency con número de intervalos igual a dos (presencia y ausencia de precipitación pluvial).
2. Aprendizaje automático: Se realizó mediante el método de aprendizaje K2 learning, con máxima verosimilitud y número máximo de padres igual a 5, sin restricciones.
3. Post-proceso:

    a) Formulación del esquema de dependencias entre variables: A partir del conjunto de datos de las 18 estaciones climatológicas, se realizó un análisis de dependencia entre todas las variables. Este análisis consiste en la cuantificación de los diferentes tipos de dependencia y ha dado como resultado la identificación de las relaciones causales existentes entre las variables.
    b) Aprendizaje estructural de las redes Bayesianas: A partir de la identificación de las relaciones causales entre variables, se han determinado las estructuras

de redes Bayesianas que mejor describen el comportamiento de la precipitación pluvial en las diferentes localidades del estado de Querétaro.

c) Aprendizaje paramétrico de la RB: Una vez obtenida la estructura topológica de la RB, se han obtenido los parámetros o distribuciones de probabilidades condicionadas entre variables que permiten representar cada uno de los arcos que componen las estructuras de RB. Estos parámetros permiten obtener las probabilidades de ocurrencia de precipitación pluvial para cualquiera de las variables que componen la red [16].

4. Resultados y discusión

Se obtuvo un modelo de RB a partir de los datos de las 18 estaciones meteorológicas del estado de Querétaro, el cual se muestra en la Fig. 4.

Fig. 4. Modelo de RB que muestra las relaciones entre las estaciones meteorológicas.

Se observa la interacción de dos localidades: Amealco y C. Sur que se estan presentes con CEA, las cuales se encuentran distantes unas de otras, lo cual sugiere un sistema pluvial con alta probabilidad con dirección Sur a Norte incluso interactuando con Sta. Rosa y Altamira que se encuentran a mayor distancia y que no incluyen otras localidades más cercanas. Otro sistema pluvial indica la interacción entre localidades orientadas de Oeste a Este como son: El Milagro, C. Civico y C. Sur con Tequisquiapan y SJR las cuales se encuentran a mayor distancia. Y un tercer sistema pluvial con dirección Oeste a Noreste que involucran las localidades El Milagro, corregidora con Colon que se encuentra a mayor distancia y que no incluyen localidades en puntos intermedios, quizás debido a la topografía de la región.

Las interacciones entre localidades implican la presencia de lluvia en estos sitios y aunque no necesariamente al mismo tiempo, debe ser en espacios de tiempo breve. Por lo que dichas interacciones pueden estar relacionadas con los vientos dominantes, los cuales tienen un efecto de movilidad sobre la precipitación pluvial. Por otra parte, la dirección de las interacciones es de localidades con climas húmedos hacia aquellas que tienen clima más seco.

Table 1. Probabilidades condicionales de las principales relaciones inferidas a partir de la RB obtenida.

CEA .77	**Colon** .875	**Sta Rsa** .9	**C.Civico** .82	**Altamira** .88	**Rochera** .84	**El Marques** .99	**SJR** .99	**Tequis-quiapan** .99
C.Sur	Huimilpan	Amealco	Amealco	Amealco	El Milagro	Huimilpan	El Milagro	El Milagro
Amealco	Milagro	CEA	Huimilpan	C.Sur	Corregi-dora	El Milagro	C.Civico	C.Civico
	Corregi-dora			CEA	C.Civico	C.Civico	El Marquez	C.Sur
						Jalpan	Colon	
							E.Montes	

Las interacciones que involucran pocas localidades tienen valores de probabilidad menores, las cuales se muestran en la Tabla 1, que aquellas en las que el número de localidades es mayor y se encuentran más distantes entre sí, lo que sugiere estar relacionado con fenómenos meteorológicos locales, en el primer caso y generalizados para el segundo caso. Esto se corrobora tomando en cuenta que las interacciones con mayor número de localidades y mayor probabilidad, también son aquellas que presentan un clima más seco, en donde la presencia de precipitación pluvial se da en eventos generalizados y no locales, bajo la influencia de otras localidades con climas más húmedos.

Por otra parte, se presentan localidades cercanas que muestran independencia en la precipitación pluvial, lo cual puede ser un factor de interés relacionado con la presencia de corrientes de aire que no permiten que la precipitación sea homogénea, a pesar de que la topografía de la región sea regular.

Para los alcances de este trabajo se cubierto con el objetivo, sin embargo, es necesario realizar un estudio más detallado tomando en cuenta otros métodos de *machine learning* que permitan realizar una comparación.

5. Conclusiones

Es posible determinar las relaciones de dependencia causal entre los valores de ocurrencia de precipitación pluvial en diferentes localidades mediante el empleo RB, aún a partir de datos incompletos. Las RB permiten el cálculo de la distribución de probabilidad condicional de la precipitación en función de diferentes localidades tanto cercanas como distantes. Estas observaciones confirman que las RB pueden ser usadas para identificar relaciones previamente indeterminadas entre variables climáticas en localidades distantes y cuantificarlas mediante inferencia con el objetivo de

realizar predicciones. Se sugiere la realización de estudios posteriores que impliquen otras variables climáticas como dirección de los vientos dominantes, presión atmosférica y temperatura del aire, que puedan ampliar la compresión de los fenómenos meteorológicos y así realizar mejores predicciones.

Referencias

1. Cano R, Sordo C, Gutierrez J.M.: Applications of Bayesian Networks in Meteorology. In Gámez et al. (eds) Advances in Bayesian Networks, Springer, pp. 309-327 (2004)
2. Cofiño AS, Cano R, Sordo C, Gutierrez J.M.: Bayesian Networks for Probabilistic Weather Prediction: Proceedings of the 15th European Conference on Artificial Intelligence, pp. 695-700 (2002)
3. Correa M, Bielza C, Paimes-Teixeira J, Alique J.R.: Comparison of Bayesian networks and artificial neural networks for quality detection in a machining process. Expert Syst. Appl., 36(3): 7270-7279 (2009)
4. Easterling D.R., Peterson T.C.: A new method for detecting undocumented discontinuities in climatological time series. Int. J. Climatol., 15(4): 369-377 (1995)
5. Gámez J.A., Mateo J.L., Puerta J.M.: Learning Bayesian networks by hill climbing: efficient methods based on progressive restriction of the neighborhood. Data Min. Knowl. Discov., 22: 106-148 (2011)
6. García, E (1988). Modifications to Köppen classification system climate to suit the conditions of the Mexican Republic. Offset Larios S.A. Mexico, pp. 201-217.
7. Garrote, L, Molina, M, Mediero, L.: Probabilistic Forecasts Using Bayesian Networks Calibrated with Deterministic Rainfall-Runoff Models. In Vasiliev et al. (eds.), Extreme Hydrological Events: New Concepts for Security, Springer, pp. 173-183 (2007)
8. Guoliang, L.: Knowledge Discovery with Bayesian Networks. PhD dissertation, National University of Singapore, Singapore (2009)
9. Guttman, N.B.: Statistical descriptors of climate. Bull. Am. Meteorol. Soc., 70: 602-607 (1989)
10. HongRui, W., LeTian, Y., XinYi, X., QiLei, F., Yan, J., Qiong, L., Qi, T.: Bayesian networks precipitation model based on hidden Markov analysis and its application. Sci China Tech. Sci., 53(2): 539-547 (2010)
11. Hruschka, E., Ebecken N.F.F.: Bayesian networks for imputation in classification Problems. J. Intell. Inform. Syst., 29: 231-252 (2007)
12. INEGI 2011 www.inegi.org.mx
13. Kazemnejad, A., Zayeri, F., Hamzah, N.A., Gharaaghaji, R., Salehi, M.: A Bayesian analysis of bivariate ordered categorical responses using a latent variable regression model: Application to diabetic retinopathy data. Sci. Res. Essays, 5(11): 1264-1273 (2010)
14. Landsberg, H.E.: Weather 'normals' and normal weather. Wkly. Weather Crop Bull., 42: 7-8 (1955)
15. Lima, C.H.R., Lall, U.: Spatial scaling in a changing climate: A hierarchical bayesian model for non-stationary multi-site annual maximum and monthly streamflow. J. Hydrol., 383(3): 307-318 (2010)
16. Mediero, O.L.: Probabilistic forecast flood flows Through Bayesian Networks Applied to a Distributed Hydrological Model. PhD dissertation, Polytechnic University of Madrid, Madrid, Spain (2007)
17. Naveed, N., Choi, M.T.S., Jaffar, A.: Malignancy and abnormality detection of mammograms using DWT features and ensembling of classifiers. Int. J. Phy. Sci., 6(8): 2107-2116 (2011)
18. Pearl, J.: Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan Kaufmann, San Mateo CA, United States, pp. 64-70 (1988)

19. Reyes, P.: Bayesian networks for setting genetic algorithm parameters used in problems of geometric constraint satisfaction. Intell. Artificial., 45: 5-8 (2010)
20. Subramaniam, T., Jalab, H.A., Taqa, A.Y.: Overview of textual antispam filtering techniques. Int. J. Phys. Sci., 5(12): 1869-1882 (2010)
21. Tae-wong, K., Hosung, A., Gunhui, C.H., Chulsang, Y.: Stochastic multi-site generation of daily rainfall occurrence in south Florida, Stoch. Environ. Res. Risk Assess., 22: 705-717 (2008)