Modelos de Mixturas de Betas en Redes Bayesianas Híbridas

Ana M. Sánchez[1] y Antonio Salmerón[1,2]

[1] Departamento de Matemáticas, Universidad de Almería, 04120 Almería, España
[2] Centro de Desarrollo y Transferencia de Investigación Matemática a la Empresa (CDTIME), Universidad de Almería, 04120 Almería, España
{asg718,antonio.salmeron}@ual.es

Abstract. Las redes bayesianas híbridas han sido estudiadas durante años por sus diversas aplicaciones. Para su modelización, varios trabajos de las últimas décadas se han basado en las MoTBFs (Mixtures of Truncated Basis Functions), que engloban tanto a las MoPs (Mixtures of Polynomials) como a las MTEs (Mixtures of Truncated Exponentials). En este trabajo, proponemos una nueva metodología para su modelización. En lugar de aproximar las funciones de densidad de cada variable mediante polinomios o exponenciales, optamos por el uso de distribuciones Beta o la combinación de éstas mediante una mixtura, dando lugar a mixturas de Betas. Estos serán denominados modelos MB o aproximaciones MB. El objetivo principal es desarrollar un modelo útil y aplicable a situaciones reales similar al modelo MoP. Para ello, es necesario llevar a cabo una serie de tareas que nos conduzcan al resultado final de la tesis. Inicialmente comprobaremos que las distribuciones Beta constituyen una herramienta adecuada para realizar este tipo de aproximaciones. A continuación, investigaremos métodos que, a partir de un conjunto de datos, permitan aprender tanto la estructura de la red bayesiana como los parámetros del modelo. Finalmente, abordaremos la inferencia probabilística sobre dichas redes con el nuevo modelo MB.
Este estudio sigue la línea de trabajos previos, pero aporta una nueva perspectiva en las aproximaciones. Nótese que la inspiración inicial surge de observar la gran variedad de formas que puede tomar una función de densidad de una distribución Beta en función de los valores de sus parámetros.

Keywords: Redes bayesianas híbridas · Distribución Beta · Inferencia probabilística.

1 Información previa sobre la tesis doctoral

Título: Modelos de Mixturas de Betas en Redes Bayesianas Híbridas.
Doctoranda: Ana M. Sánchez González.
Correo de contacto: asg718@ual.es.
Director y tutor: Antonio Salmerón.
Universidad: Universidad de Almería.

Departamento: Departamento de Matemáticas.
Línea de Investigación: Estadística e Investigación Operativa.
Fecha de inicio: 19 de noviembre de 2024.

2 Introducción

En torno a la década de 1990, las redes bayesianas [1] surgieron como una herramienta estadística potente dentro del ámbito del machine learning. Una de sus grandes ventajas es la incorporación de las relaciones de independencia condicional entre las variables mediante un grafo acíclico dirigido. En él, los nodos representan las variables y las aristas indican el tipo de relaciones entre ellas.

Actualmente, las redes bayesianas se utilizan para una gran variedad de modelos y en ámbitos muy diversos como pueden ser desde la economía a la medicina, desde el análisis de riesgos en finanzas al diagnóstico médico en función de los síntomas, entre muchas otras [2].

3 Investigaciones previas

Debido a la naturaleza de estos modelos, en numerosas ocasiones, las redes bayesianas deben contener tanto variables continuas como discretas. En este caso, son denominadas *redes bayesianas híbridas*.

En los años 2000, uno de los modelos implementados fue el basado en distribuciones condicionales lineales gaussianas o el modelo de mixturas de distribuciones gaussianas [3], incluyendo modelos con variables continuas y discretas.

Posteriormente, surgieron otras formas de abordarlas mediante el modelo MoP [4] y el MTE [5]. En 2012, ambos fueron generalizados mediante los modelos MoTBFs [6].

La motivación para realizar esta tesis surge a partir de estos trabajos. La idea principal es basarse en la distribución Beta, debido a la gran diversidad de formas que puede adoptar en función de solo dos parámetros. Con ella se pretende modelar las variables continuas de las redes bayesianas híbridas mediante mixturas de Betas de una forma más compacta que con MoTBFs.

4 Objetivos

Los objetivos de la tesis son:

- Inicialmente, realizar un revisión bibliográfica en este ámbito.
- A continuación, comprobar teóricamente y prácticamente la aplicabilidad y el funcionamiento de las distribuciones Beta en el modelado de redes bayesianas híbridas.

– Hallar métodos de aprendizaje tanto para la estructura de la red como para los parámetros de las distribuciones a partir de los datos, una vez conocida la estructura del grafo.
– Finalmente, desarrollar procedimientos para realizar inferencia probabilística con dicho modelo.

5 Metodología y planificación

5.1 Metodología

En primer lugar, se realiza un estudio previo acerca de todos los aspectos anteriores presentes en modelos ya existentes, como por ejemplo las MoTBFs, para dar paso a modelos que involucren las mixturas de Betas. El problema será abordado desde la perspectiva bayesiana y los avances metodológicos se contrastarán experimentalmente. Además, todas las implementaciones serán llevadas a cabo en R.

5.2 Planificación

A continuación, se expone la planificación inicial para el proyecto de esta tesis a tiempo completo que se realiza a la vez que el trabajo en el puesto de profesora sustituta en la Universidad de Almería.

Primer año: Los primeros meses de la tesis han estado centrados en realizar una revisión bibliográfica tanto sobre las redes bayesianas como el uso de las MoTBFs en ellas. En particular, las MTEs y las MoPs, haciendo especial hincapíe en estas últimas. A continuación, se realizan experimentos iniciales que contrastan la viabilidad de las distribuciones Beta en el modelado de dichas redes. Para ello se utiliza R y el paquete MoTBF [7].

Segundo año: Se continúa con los experimentos a través de mixturas de Betas en redes Bayesianas híbridas ampliando los realizados en el primer año. Se comienza a redactar un primer artículo. Avanzamos al siguiente objetivo de la tesis: desarrollar otros métodos de aprendizaje para los parámetros de los potenciales de las redes a partir de los datos y dada la estructura de la red.

Tercer año: Finalización del apartado de aprendizaje de los parámetros para las aproximaciones de las funciones de cada variable de la red. Se inicia la búsqueda de métodos para obtener la estructura de la red o el grafo, dados los datos. Todo ello se comprobará con experimentos en R o Phyton, aglutinando todas las metodologías desarrolladas.

Cuarto año: El último año estará dedicado al apartado de inferencia, a la finalización y revisión de los experimentos, así como al comienzo y término de la redacción del documento final de la tesis.

6 Algunos resultados

Una de las comprobaciones llevadas a cabo ha sido la evaluación de las distribuciones Beta y sus respectivas mixturas en la red bayesiana híbrida expuesta a continuación. A partir del 80% del conjunto de datos del archivo "ecoli", usados en [7], y la estructura de la red dada por la Figura 1, las variables discretas se modelan mediante distribuciones Bernoulli y las continuas con mixturas de Betas, o en particular, una única Beta. Para poner de manifiesto las relaciones de dependencia entre las variables, se utiliza la herramienta de discretización de los respectivos padres en cada caso. Además, respecto a las variables discretas dicotómicas lip y chg, los valores correspondientes para cada una se representan mediante 0 y 1. El resto de variables de la red son continuas y en nuestro caso debemos asegurarnos, antes de calcular las aproximaciones, de que sus respectivos valores están comprendidos entre 0 y 1, puesto que es un requisito del modelo MB. En caso contrario, se aplicará una transformación lineal que permita obtener el dominio deseado. Dicha transformación no afecta a la especificación de las distribuciones MoTBF, dado que es reversible y no altera el conjunto de funciones base usado [8]. Para que la comparación sea equitativa, ambos modelos serán aprendidos con los datos transformados en caso de que así se requiera.

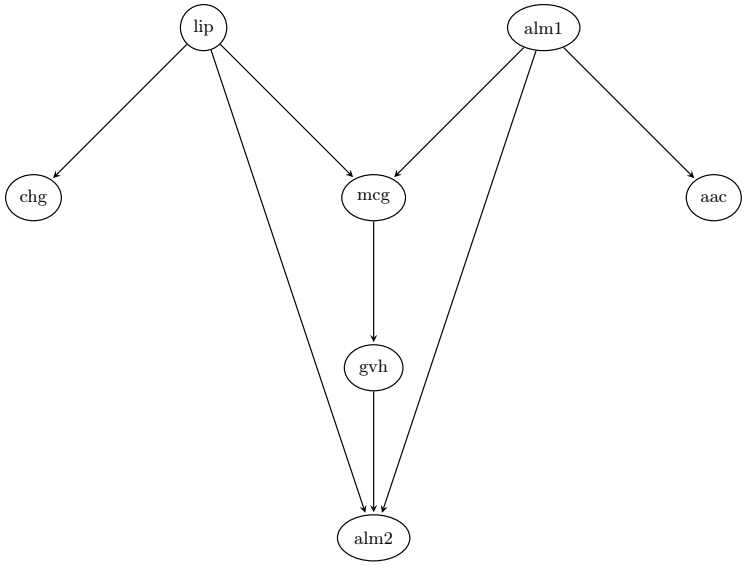


Fig. 1. Red bayesiana híbrida.

En el cálculo de las aproximaciones MoPs se usa el paquete MoTBF en las mismas condiciones que en [7]: máximo 4 particiones por variable y en cada una, un máximo 12 parámetros por intervalo de partición.

En el modelo MB, las variables discretas se modelan con distribuciones de Bernoulli, mientras que las continuas son ajustadas mediante una Beta con el paquete *EnvStats* o una mixtura de Betas gracias al paquete *BetaMixture* en cada una de las particiones. De este modo, para la variable alm1 se utiliza una mixtura de Betas con 4 componentes; para la distribución de aac | alm1, 2 particiones, cada una con una mixtura de Betas con 2 componentes; para mcg | { lip, alm1}, 4 particiones; para gvh | mcg, 2 particiones; y para alm2 | {lip, gvh, alm1}, 4 particiones.

En la Figura 2, se muestra un ejemplo gráfico de las aproximaciones obtenidas por ambos modelos para la función de densidad de la variable continua alm1. Nótese cómo la mixtura de betas ofrece un mejor ajuste, al ser más estable en las colas de la distribución. En ambos casos, el número de parámetros es igual a 12.

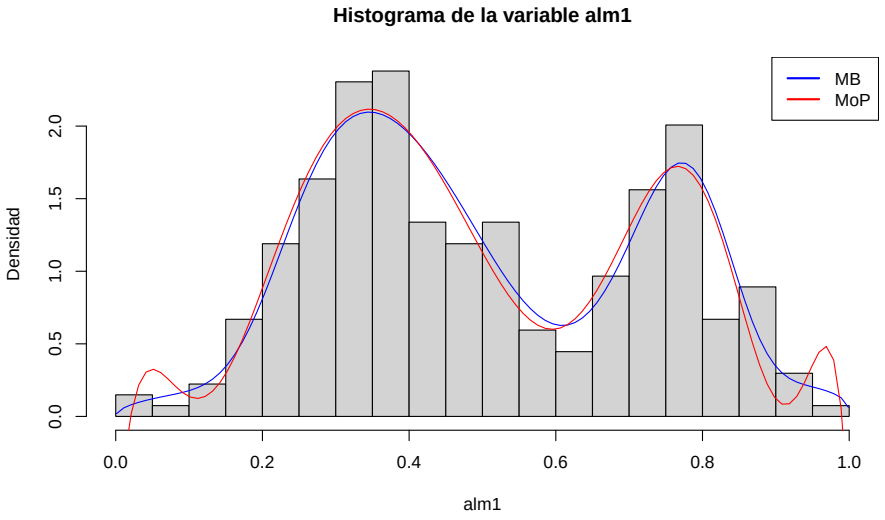


Fig. 2. Representación de las aproximaciones obtenidas por ambos modelos.

A partir de la Regla de la Cadena se obtienen las funciones de probabilidad conjuntas del modelo MoP y del modelo propuesto de Betas, MB.

Por último, estudiamos la log-verosimilitud y el BIC de ambos modelos, obteniendo los resultados que aparecen en la Tabla 1, en la que se puede observar el hecho de que para log-verosimilitudes muy similares, el menor número de parámetros requerido implica que el BIC en el modelo MB sea superior al del modelo MoP. Ésta es una ventaja que se observa sucesivamente en el resto de experimentos. En general, el modelo MB requiere menos parámetros para obtener aproximaciones tan satisfactorias como las del modelo MoP.

Table 1. Logverosimilitud y BIC del modelo MoP y del MB para la red bayesiana de la Figura 1.

Modelo	Log-verosimilitud	BIC	Número de parámetros
MoP	762.44	398.78	130
MB	775.16	498.23	99

7 Importancia

En este trabajo, se expone un plan de investigación centrado en la mejora de modelos de redes bayesianas híbridas, las cuales, tal y como se menciona al inicio de este documento, poseen múltiples aplicaciones en la vida real.

Como se ha puesto de manifiesto en el ejemplo del apartado anterior, gracias a una serie de experimentos, se ha comprobado la eficacia del modelo MB. Una de sus principales ventajas, por el momento, es el menor número de parámetros requerido en comparación con el modelo MoP.

Además, en el estudio teórico se comprueba que las distribuciones conjuntas entre variables continuas y discretas mantienen ciertas estructuras similares, al igual que en sus respectivas marginalizaciones, lo que supone una ventaja a la hora de implementar procesos de inferencia probabilística.

Acknowledgments. (Agradecimientos.) Este trabajo está financiado por MICIU/AEI/ 10.13039/501100011033 y ERDF "A way of making Europe" (Grant PID2022-139293NB-C31).

Disclosure of Interests. (Declaración de Ausencia de Conflicto de Intereses.) Los autores no se encuentran en una situación de conflicto de intereses que sea relevante en el contexto de este artículo.

References

1. Pearl, J.: Probabilistic reasoning in intelligent systems. Morgan-Kaufmann (San Mateo). 1988.
2. Salmerón, A.: Redes bayesianas como modelos generativos: de los juegos a las finanzas. En: La Inteligencia Artificial hoy y sus aplicaciones con Big Data. Funcas, pp. 145-164. 2025
3. Shenoy, P. P.: Inference in hybrid Bayesian networks using mixtures of Gaussians. In: Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence, pp. 428–436. AUAI Press, Cambridge, MA, USA (2006)
4. Shenoy, P. P. , West, J. C. : Inference in hybrid Bayesian networks using mixtures of polynomials. International Journal of Approximate Reasoning **52**(5), 641–657 (2011)
5. Moral, S., Rumí, R., Salmerón, A.: Mixtures of Truncated Exponentials in Hybrid Bayesian Networks. In: Benferhat, S., Besnard, P. (eds.) Symbolic and Quantitative Approaches to Reasoning with Uncertainty. ECSQARU 2001. Lecture Notes in Computer Science(), vol. 2143, pp. 156–167. Springer, Heidelberg (2001). `https://doi.org/10.1007/3-540-44652-4_15`

6. Langseth, H., Nielsen, T. D., Rumí, R., Salmerón, A. : Mixtures of truncated basis functions. International Journal of Approximate Reasoning **53**(2), 212–227 (2012)
7. Pérez-Bernabé, I., Maldonado, A. D., Nielsen, T. D., Salmerón, A. : MoTBFs: An R package for learning hybrid bayesian networks using mixtures of truncated basis functions. R Journal **12**, 342–358 (2020)
8. Salmerón, A., Langseth, H., Masegosa, A.R., Nielsen, T.D.: A reparameterization of mixtures of truncated basis functions and its applications. Proceedings of The 11th International Conference on Probabilistic Graphical Models, PMLR 186:205-216. 2022.