Resumen del Articulo "Decision Boundary for Discrete Bayesian Network Classifiers"

por Gherardo Varando, Concha Bielza y Pedro Larrañaga

Gherardo Varando

1. Introducción

El articulo "Decision Boundary for Discrete Bayesian Network Classifier" [9] ha sido aceptado para la publicación en el Journal of Machine Learning Research (IF de 2.853). En ese trabajo hemos analizado la capacidad expresiva de los clasificadores binarios basados en redes Bayesianas formadas por variables predictoras categóricas. Un clasificador binario es un algoritmo que a partir de una base de datos de entrenamiento, formada por n variables aleatorias $X_1, \dots, X_n$ llamadas predictoras y una variable binaria llamada clase $C \in \{-1, +1\}$, intenta aprender una función que *clasifica* nuevas instancias. El objetivo es entonces aprender una función de decisión,

$$f : \Omega = \Omega_1 \times \cdots \times \Omega_n \mapsto \{-1, +1\}.$$

En este trabajo, concentramos nuestra atención sobre clasificadores Bayesianos y en particular clasificadores basados en redes Bayesianas. Los clasificadores con redes Bayesianas estiman a partir de los datos de entrenamiento la distribución conjunta de las variables $X_1, \dots, X_n, C$, $P(C, X_1, \dots, X_n)$, y en particular dado que se asume una estructura de redes Bayesianas, la distribución de probabilidad P se puede factorizar como,

$$P(C, X_1, \dots, X_n) = P(C) \prod_{i=1}^{n} P(X_i | \mathbf{X}_{\mathbf{pa}(i)}),$$

donde $\mathbf{X}_{\mathbf{pa}(i)}$ indica los padres de X_i en la red Bayesiana (nodos con arcos dirigidos a X_i). A partir de la distribución conjunta es posible definir la función de decisión generada como,

$$\begin{aligned} f(x_1, \dots, x_n) &= \arg \max_{c \in \{-1,+1\}} P(C = c | X_1 = x_1, \dots, X_n = x_n) \\ &= \arg \max_{c \in \{-1,+1\}} P(C = c, X_1 = x_1, \dots, X_n = x_n) \\ &= \arg \max_{c \in \{-1,+1\}} P(C = c) \prod_{i=1}^{n} P(X_i = x_i | \mathbf{X}_{\mathbf{pa}(i)}) \end{aligned}$$

Es decir, para cada nueva instancia de los predictores, $(x_1, \dots, x_n)$, el clasificador predice la clase mas máxima probabilidad a posteriori.

En nuestro trabajo hemos analizando la expresividad de los clasificadores basados en redes Bayesianas y variables predictoras categóricas, es decir cuantas y cuales entre todas las posibles funciones de decisión se pueden generar con clasificadores basado en redes Bayesianas. Hemos estudiado particular clasificadores *Bayesian network-augmented naive Bayes* (BAN), un tipo de

clasificadores que son una directa extensión del modelo naive Bayes (NB), donde se permite la estructura de una rede Bayesiana cualquiera entre las variables predictoras, manteniendo la variable clase como padre de todas las predictoras, esto último exactamente como en el naive Bayes.

2. Estado del Arte

El primer resultado sobre la capacidad expresiva de clasificadores con redes Bayesianas fue demostrado por Marvin Minsky (premio Turing en 1969 y premio Fundación BBVA Fronteras del Conocimiento en 2013) en 1961 [5], que probó que un clasificador naive Bayes con predictoras binarias solo puede generar funciones de decisión lineales.

Peot [7] en 1996 revisó el resultado de Minsky sobre naive Bayes y presentó algunas ampliaciones. En particular analizó predictoras categóricas, todas con el mismo numero de valores, y el caso de dependencias entre predictoras. Ademas, Peot [7] analizó cotas superiores del numero de funciones de decisión que puede representar el naive Bayes.

Domingos y Pazzani [2] estudiaron a fondo la optimalidad del clasificador naive Bayes. Demostraron que el clasificador naive Bayes puede conseguir alcanzar el error optimo también si las asunciones de independencias no se cumplen.

Jaeger [3] demostró que, para predictores binarios, la expresividad de los clasificadores con redes Bayesianas de diferentes niveles de complejidad está relacionada con polinomios de diferentes grados.

Ling y Zhang [4] probaron que un clasificador con redes Bayesianas donde cada nodo tiene como maximo k padres no puede representar una función de decisión que contenga $(k+1)$-XOR.

Nakamura *at al.* [6] investigaron el espacio con producto interno asociado con un clasificador con redes Bayesianas, es decir, el mínimo espacio vectorial con producto escalar tal que representa las funciones de decisión generadas. Probaron cotas inferiores y superiores para la dimensión del espacio con producto interno y enlazaron la dimensión del espacio con producto interno asociado con la dimensión de Vapnik-Chervonenkis (VC) [8].

Yang y Wu [10] demostraron que en el caso de redes Bayesianas completas o en general sin V-estructuras, la dimensión del *inner product space* es igual a la dimension de Vapnik-Chervonenkis del conjunto de funciones de decisiones generadas.

3. Resumen de los Resultados

Nuestros resultados son una ampliación de los resultados de Minsky [5] y Peot [7], y están basados en una representación polinomial del logaritmo de la función de probabilidad y de probabilidad condicionada de una variable aleatoria categórica. En particular si $X_1, X_2, \ldots, X_n$ son variables categóricas que toman valores respectivamente sobre los conjuntos $\Omega_1, \Omega_2, \ldots, \Omega_n$, se obtiene que

$$P\left(X_1 = x_1 | X_2 = x_2, \ldots, X_n = x_n\right) = \prod_{(j_1,\ldots,j_n)} p(j_1|j_2,\ldots,j_n)^{\prod_{i=1}^{m} \ell_{j_i}^{\Omega_i}(x_i)}, \tag{1}$$

donde $\ell_{j_i}^{\Omega_i}(x_i)$ es el $j_i$ésimo elemento de la base polinomial de Lagrange [1] sobre los puntos de Ω_i. Gracias a esa representación, hemos demostrado el siguiente resultado válido para naive Bayes:

Teorema 1 *Sea f una función de decisión para un problema de clasificación binaria sobre n variables categóricas $X_i \in \Omega_i = \{\xi_i^1, \ldots, \xi_i^{m_i}\}$, donde $|\Omega_i| = m_i$. Entonces f se representa por el signo de un polinomio del tipo $\sum_{i=1}^{n} \left(\sum_{j=1}^{m_i} \alpha_i(j) \ell_j^{\Omega_i}(x_i) \right)$ si y solo si existe un clasificador naive Bayes tal que genera f, donde $\ell_j^{\Omega_i}$ son los polinomios de la base de Lagrange sobre Ω_i.*

El Teorema 1 implica que el siguiente espacio de polinomios es suficiente para representar todas las funciones de decisión que se pueden generar con un clasificador naive Bayes,

$$\mathcal{P}^{NB} = \left\{ \sum_{i=1}^{n} \left(\sum_{j=1}^{m_i} \alpha_i(j) \ell_j^{\Omega_i}(x_i) \right) \text{ t.q. } \alpha_i(j) \in \mathbb{R} \right\}.$$

Además, todas las funciones de decisión representadas por el signo de un polinomio en $\mathcal{P}^{NB}$ se pueden generar por un clasificador naive Bayes. La demostración que hemos desarrollado es constructiva y explicamos cómo construir el polinomio cuyo signo representa la función de decisión generada. Y recíprocamente como, a partir de los coeficientes de un polinomio en la forma específica es posible definir probabilidades sobre una estructura de naive Bayes tales que la función de decisión generada es igual al signo del polinomio.

Más aún el resultado para clasificadores naive Bayes se puede demostrar en general para clasificadores BAN, siempre que en el sub-grafo de las variables predictoras no haya V-estructuras.

Teorema 10 *Sea $\mathcal{G}$ un grafo acíclico dirigido, con nodos X_i por cada $i = 1, \ldots, n$, y sea f una función de decisión con predictoras $X_i \in \Omega_i = \{\xi_i^1, \ldots, \xi_i^{m_i}\}$. Supongamos, además que $\mathcal{G}$ no contiene V-estructuras, entonces f se puede representar con el signo del siguiente polinomio*

$$r(\mathbf{x}) = \sum_{i=1}^{n} \sum_{j=1}^{m_i} \ell_j^{\Omega_i}(x_i) \sum_{\mathbf{k} \in \mathbb{M}_i} \beta_i(j|\mathbf{k}) \prod_{s \in \mathbf{pa}(i)} \ell_{k_s}^{\Omega_s}(x_s),$$

si y solo si f es generada por un clasificador BAN con sub-grafo de las predictoras dado por $\mathcal{G}$.

En el caso de clasificadores BAN sin V-estructuras se obtiene que el siguiente espacio de polinomios representa las funciones de decisión generadas por clasificadores BAN con grafo de las predictoras $\mathcal{G}$,

$$\mathcal{P}_{\mathcal{G}}^{BAN} = \left\{ r(\mathbf{x}) = \sum_{i=1}^{n} \sum_{j=1}^{m_i} \ell_j^{\Omega_i}(x_i) \sum_{\mathbf{k} \in \mathbb{M}_i} \beta_i(j|\mathbf{k}) \prod_{s \in \mathbf{pa}(i)} \ell_{k_s}^{\Omega_s}(x_s) \text{ t.q. } \beta_i(j|\mathbf{k}) \in \mathbb{R} \right\}.$$

En caso de presencia de V-estructuras una de las implicaciones del Teorema 10 es válida. En particular para cada función de decisión generada por un clasificador BAN existe un polinomio de la forma descrita cuyo signo es igual a la función de decisión. Recíprocamente en presencia de V-estructuras no siempre es verdad que cada polinomio de esa forma represente una función de decisión relacionada con un clasificador BAN con esa estructura.

Los espacios de polinomios $\mathcal{P}^{NB}$ y en general $\mathcal{P}_{\mathcal{G}}^{BAN}$ son espacios vectoriales que podemos analizar como sub-espacios del siguiente conjunto:

$$\mathcal{P}^{FBN} = \left\{ \sum_{\mathbf{k} \in \mathbb{M}} \gamma_{\mathbf{k}} \prod_{i=1}^{n} \ell_{k_i}^{\Omega_i}(x_i) \text{ t.q. } \gamma_{\mathbf{k}} \in \mathbb{R} \right\},$$

donde la suma se extiende sobre todos los posibles multi-indices $\mathbf{k}$. El espacio $\mathcal{P}^{FBN}$, que equivale a $\mathcal{P}_{\mathcal{G}}^{BAN}$ si $\mathcal{G}$ es una rede Bayesiana completa, es un espacio vectorial de dimensión $|\Omega| = \prod_{i=1}^{n} |\Omega_i|$ y contiene todos los polinomios cuyo grado en la $i$ésima variable es minore o igual que $|\Omega_i| - 1$. Es claro por la definición que los polinomios en $\mathcal{P}^{FBN}$ pueden interpolar cualquiera función sobre Ω y entonces pueden representar cualquier clasificador sobre las variables $X_1, \ldots, X_n$.

En el articulo hemos computado la dimensión de los espacios $\mathcal{P}^{NB}$ y $\mathcal{P}_{\mathcal{G}}^{BAN}$ y gracias a eso hemos demostrado una cota superior para el número de funciones de decisión generadas por un clasificador BAN. Concretamente el siguiente resultado.

Corolario 18 *Se considera una clasificador BAN sobre variables predictoras* $X_i \in \Omega_i$, $|\Omega_i| = m_i$ *y supongamos que el sub-grafo de las predictoras* $\mathcal{G}$ *no contiene* V*-estructuras. Entonces*

$$2^d \leq |signo(\mathcal{P}_{\mathcal{G}}^{BAN})| \leq 2\sum_{k=0}^{d-1} \binom{M-1}{k},$$

donde $d = \sum_{i=1}^{n} \left((m_i - 1)\prod_{s\in \mathbf{pa}(i)} m_s\right) + 1$ *y* $M = \prod_{i=1}^{n} m_i$.

En resumen en el artículo hemos realizado desarrollos metodológicos para analizar las funciones de decisión y el poder expresivo relacionado con clasificadores BAN. Se ha demostrado cómo la complejidad y la estructura de un clasificador BAN esta relacionada con la complejidad expresiva de las funciones de decisión que puede generar. En ausencia de V-estructuras los resultados permiten acotar la capacidad expresiva de cada estructura y asociar a cada familia de clasificadores un espacio de polinomios que los caracteriza. Entre las posibles líneas futuras se podrían extender las propriedad geométricas de los espacios polinomiales definidos y desarrollar algoritmos de aprendizaje de clasificadores más eficientes.

Referencias

[1] Milton Abramowitz and Irene A. Stegun. *Handbook of Mathematical Functions: With Formulas, Graphs, and Mathematical Tables*. Applied Mathematics Series. Dover Publications, 1964.

[2] Pedro Domingos and Michael Pazzani. On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning*, 29(2-3):103–130, 1997.

[3] Manfred Jaeger. Probabilistic classifiers and the concepts they recognize. In Tom Fawcett and Nina Mishra, editors, *Proceedings of the Twentieth International Conference on Machine Learning (ICML-03)*, pages 266–273. AAAI Press, 2003.

[4] Charles X. Ling and Huajie Zhang. The representational power of discrete Bayesian networks. *Journal of Machine Learning Research*, 3:709–721, 2002.

[5] Marvin Minsky. Steps toward artificial intelligence. In *Computers and Thought*, pages 406–450. McGraw-Hill, 1961.

[6] Atsuyoshi Nakamura, Michael Schmitt, Niels Schmitt, and Hans Ulrich Simon. Inner product spaces for Bayesian networks. *Journal of Machine Learning Research*, 6:1383–1403, 2005.

[7] Mark A. Peot. Geometric implications of the naive Bayes assumption. In *Proceedings of the Twelfth International Conference on Uncertainty in Artificial Intelligence*, UAI'96, pages 414–419, San Francisco, 1996. Morgan Kaufmann Publishers Inc.

[8] Vladimir N. Vapnik and Alexy Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and Its Applications*, 16(2):264–280, 1971.

[9] Gherardo Varando, Concha Bielza, and Pedro Larrañaga. Decision boundary for discrete Bayesian network classifiers. *Journal of Machine learning Research*, 2015.

[10] Youlong Yang and Yan Wu. On the properties of concept classes induced by multivalued Bayesian networks. *Information Sciences*, 184(1):155–165, 2012.