Modelos gráficos probabilísticos para la clasificación supervisada empleando la estimación basada en kernels Gaussianos esféricos

Aritz Pérez, Pedro Larrañaga, Iñaki Inza

Intelligent Systems Group

Dept. de C.C.I.A.

Universidad del País Vasco

aritz@si.ehu.es, ccplamup@si.ehu.es, inza@si.ehu.es

Resumen

El clasificador naive Bayes ha demostrado comportarse sorprendentemente bien en la clasificación supervisada a pesar de que asume que las variables predictoras son condicionalmente independientes dada la clase, lo que generalmente no se cumple. El clasificador *red Bayesiana aumentada a árbol* rompe con esta suposición tan fuerte ya que permite dependencias entre las variables predictoras, por lo que se comporta mejor que el naive Bayes en ciertos dominios.

Muchos de los clasificadores basados en redes Bayesianas (naive Bayes, red Bayesiana aumentada a árbol, red Bayesiana k-dependiente, semi naive Bayes...) únicamente emplean variables discretas, a pesar de que muchos dominios reales incluyen variables continuas. Existen tres opciones para estimar las funciones de densidad de las variables continuas: 1.Discretizar las variables continuas con la consecuente pérdida de información. 2.Aproximarse a la función de densidad de los datos mediante una estimación paramétrica (habitualmente Gaussiana), con el consecuente error en la estimación si la distribución real difiere de la distribución paramétrica seleccionada. 3.Aproximar la densidad mediante una estimación no paramétrica (kernels,...). La estimación no paramétrica es más flexible que la estimación paramétrica, ya que se ajusta razonablemente mejor a la mayoría de las funciones de densidad.

Este trabajo presenta el paradigma *red flexible condicionada*. No pretende ser un estudio en profundidad del nuevo paradigma, sino su introducción para la clasificación supervisada. Dicho paradigma emplea la estimación basada en kernels para modelar la densidad de las variables continuas. La red flexible condicionada puede ser entendida como una extensión de los paradigmas *red Bayesiana* y *red Gaussiana condicionada*, ya que permite una estimación más flexible y precisa de la función de densidad de las variables. A modo de ejemplo práctico, se incluye la adaptación del algoritmo *red Bayesiana aumentada a árbol* de Friedman y col. (1997) a las redes flexibles condicionadas. Esta adaptación, puede ser considerada como la extensión del clasificador *flexible Bayes* de John y Langley (1995), de la misma manera que la red Bayesiana aumentada a árbol es una extensión del naive Bayes.

Además, y con el fin de sentar las bases de nuestra línea de trabajo se propone un estimador para la cantidad de información mutua entre dos variables continuas multidimensionales cuya densidad está basada en kernels.

1. Motivación

La clasificación supervisada es una tarea básica dentro del análisis de datos y el reconocimiento de patrones que requiere de la construcción de un clasificador: función que asigna una clase a una instancia descrita por un conjunto de variables.

Se han empleado numerosos paradigmas

Actas del III Taller Nacional de Minería de Datos y Aprendizaje, TAMIDA2005, pp.125-134
ISBN: 84-9732-449-8

para realizar tareas de clasificación supervisada, entre los cuales los *modelos gráficos probabilísticos* (*PGM*, *probabilistic graphical models*) [12] son uno de los más efectivos y conocidos en dominios con incertidumbre. Un PGM es un grafo acíclico dirigido con un conjunto de nodos que representan a las variables y un conjunto de arcos que representan las relaciones de (in)dependencia condicional entre las variables. Los PGMs se utilizan para codificar la distribución de probabilidad conjunta, que viene determinada por las relaciones de dependencia condicional representadas por la estructura del grafo. Esto, combinado con la regla de Bayes, puede ser empleado para clasificar. Para inducir un clasificador a partir de una base de datos, se consideran dos tipos de variables: la variable clase o *clase* C, y el resto de variables o *predictoras*, $\boldsymbol{X} = (X_1, \ldots, X_d, X_{d+1}, \ldots, X_l)$. Asumimos que $\{X_1, \ldots, X_d\}$ es el conjunto de predictoras con valores numéricos continuos y $\{X_{d+1}, \ldots, X_l\}$ es el conjunto de predictoras discretas.

El proceso de clasificación de una instancia $\boldsymbol{x}$ consiste en seleccionar la clase c con la máxima probabilidad *a posteriori*, $p(c|\boldsymbol{x})$. El proceso de clasificación empleando PGMs se puede realizar de la siguiente manera:

$$p(c|\boldsymbol{x}) \propto \rho(c, \boldsymbol{x}) = p(c) f(\boldsymbol{x}|c)$$
$$= p(c) \prod_{i=1}^{d} f(x_i|\boldsymbol{pa}_i) \prod_{j=d+1}^{l} p(x_j|\boldsymbol{pa}_j) \qquad (1)$$

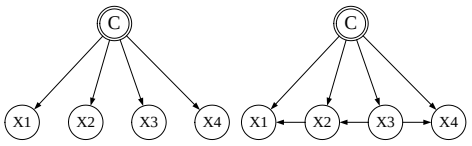
donde $\boldsymbol{pa}_i$ denota una realización de las variables $\boldsymbol{Pa}_i$, que son el conjunto de variables padres de la variable X_i. $p(\cdot)$ denota una distribución de probabilidad, $f(\cdot)$ una función de densidad y $\rho(\cdot)$ una función de probabilidad generalizada [3].

Resumiendo, un clasificador basado en el paradigma de los modelos gráficos probabilísticos viene determinado por:

1. La estructura que especifica el conjunto de relaciones de (in)dependencia condicional que se dan entre las variables, y que a su vez define una factorización concreta de la función conjunta $\rho(c, \boldsymbol{x})$.
2. El estimador empleado para modelar la función de densidad de las variables.

Desde el punto de vista del tipo de estructura uno de los clasificadores más simples y más antiguos, basado en los PGMs, es el *naive Bayes* (*NB*) [6]. El clasificador NB asume en su estructura que las variables predictoras son condicionalmente independientes entre sí dada la variable clase. La Figura 1(a) representa una estructura NB. A pesar de ello, su rendimiento es sorprendentemente bueno, incluso en las bases de datos que no siguen su suposición [5].

El buen rendimiento del clasificador NB ha motivado la investigación de paradigmas basados en PGMs que relajen la fuerte suposición de independencia. Uno de los primeros y que obtienen mejor rendimiento es la *red Bayesiana aumentada a árbol* (*TAN*, tree augmented Bayesian network) [8]. TAN construye un árbol de dependencias entre las predictoras que son a su vez hijas de la variable clase. La Figura 1(b) representa una estructura TAN.



(a) NB. (b) TAN.

Figura 1: Ejemplos de estructuras basadas en PGMS.

Para estimar la función de densidad que sigue una variable aleatoria existen tres opciones:

1. Discretizar la variable y estimar su distribución de probabilidad mediante una distribución multinomial empleando para ello los datos discretizados.
2. Estimar la función de densidad de forma paramétrica (Gaussiana,...)
3. Estimar la función de densidad de forma no paramétrica (kernels,...)

La opción más extendida dentro de los PGMs es la estimación empleando la distribución multinomial sobre la discretización de

las variables continuas. Un PGM que asume que todas las variables aleatorias siguen una distribución multinomial se conoce como *red Bayesiana* (*BN*, *Bayesian network*) [14]. Este paradigma únicamente puede manejar directamente variables discretas, y por tanto, las variables continuas deben ser previamente discretizadas, con la consecuente pérdida de información. A pesar de ello, la estimación puede ajustarse relativamente bien a funciones de densidad que no posean excesivas fluctuaciones: el número de intervalos necesarios para conseguir un buen ajuste debe crecer con el número de fluctuaciones. Esto hace que aumente el número de parámetros necesarios y disminuya el número de casos disponibles para computarlos, y por tanto disminuya a su vez la robustez de dicha estimación. En la literatura existe una amplia batería de clasificadores basados en este paradigma, que es el más empleado entre los que aquí se exponen.

La segunda opción consiste en ajustar la densidad subyacente a los datos por medio de una distribución paramétrica, más concretamente mediante la función de densidad Gaussiana o normal. Mediante esta opción se asume que las variables aleatorias continuas condicionadas a un conjunto de variables padres siguen una distribución Gaussiana condicionada. El paradigma basado en los PGMS que realiza esta suposición es conocido como *red Gaussiana condicionada* (*CGN*, *conditional Gaussian network*) [15]. Esta suposición es tremendamente exigente ya que, pese a que en dominios reales muchas variables siguen una distribución Gaussiana, otras variables pueden seguir funciones de densidad muy alejadas de la normal. Los clasificadores basados en este paradigma tienen un comportamiento comparable al de las redes Bayesianas, siempre y cuando las variables del dominio no posean una densidad muy alejada de la normal. La suposición de que las variables sigan una función de densidad condicional Gaussiana permite modelar grafos mucho más complejos que las BN, ya que un grafo completo con variables predictoras continuas y la clase, únicamente requiere de $\mathcal{O}(|C|n^2)$ parámetros para ser modelado, donde $|C|$ es la cardinalidad de la clase. Además, la estimación de los parámetros necesarios resulta más robusta que en las BN, ya que solo es necesario realizar $|C|$ particiones del conjunto de casos de entrenamiento.

El objetivo de este trabajo consiste en definir el paradigma de las *redes flexibles condicionadas* (*CFN*, *conditional flexible networks*). Una CFN ajusta la densidad de las variables continuas condicionadas a sus padres mediante una estimación basada en kernels, introducidos por Rosenblatt [16] en su forma univariada. Consideramos este paradigma como una generalización de las BN y RCG desde el punto de vista de la estimación de las densidades, ya que mediante la estimación basada en kernels se dota a las CFNs de una mayor capacidad de ajuste a las densidades reales que siguen las variables aleatorias.

Este trabajo no pretende ser un estudio en profundidad de las CFNs sino una introducción de las CFNs restringida a la clasificación supervisada. Con el objetivo de ilustrar la adaptación de algoritmos de inducción de clasificadores basados en BN y RCG el trabajo incluye la adecuación del algoritmo TAN a las CFNs, al que llamaremos *red flexible condicionada aumentada a árbol* (*TAF*, *tree augmented conditional flexible network*). TAF supone una extensión desde el punto de vista de la flexibilidad de la estimación de densidades, de los clasificadores con estructura TAN basados en las BNs y RCGs. A su vez y desde el punto de vista de la complejidad de las dependencias permitidas, TAF es una extensión del clasificador flexible Bayes [1] (*FB*) [10], ya que se rompe con la suposición de independencia condicional dada la clase realizada por NB. Esta idea se encuentra ilustrada en la Figura 2.

Este trabajo se centra en las dificultades que surgen en la modelización que se hace de las variables continuas y sus relaciones en el paradigma de la *red flexible condicionada* (*CFN*, conditional flexible network). A lo largo del trabajo nos centraremos en modelos con variables predictoras exclusivamente

[1] El clasificador flexible Bayes equivale a un *Parzen Window Classifier* con un kernel multidimensional con $H = diag(s_1, ..., s_d)$ o con su equivalente kernel producto multidimensional

continuas. El tratamiento adicional de las variables discretas no supone una dificultad añadida ya que es análoga a la modelización que se hace de los dominios mixtos en las *redes condicionales Gaussianas* [15]: condicionar una densidad a una variable discreta equivale a crear particiones de la base de datos en función de sus valores y aprender una función de densidad en cada una de las particiones. Cada una de las partes se corresponde con una función de densidad similar a la original, pero que es modelada únicamente a partir de la partición que le corresponde.

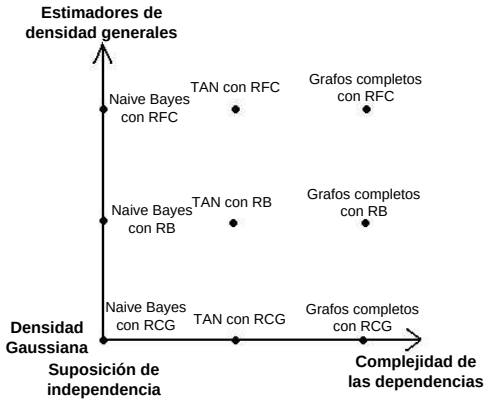


Figura 2: Representación de diferentes PGMs en función de la complejidad de las dependencias y de la estimación de las densidades reales

El resto del trabajo se organiza de la siguiente manera: en la Sección 2 se introduce el estimador no paramétrico basado en kernels Gaussianos esféricos que se empleará para modelar las CFNs. En la misma sección se mencionan los pros y los contras de este estimador frente a otros. La Sección 3 describe el paradigma CFN y se centra en el problema de seleccionar buenos parámetros de suavizado para su modelización. En la Sección 4 se adapta el algoritmo TAN a las CFN y se comentan sus características computacionales. Para ello se propone un estimador para la cantidad de información mutua y de cantidad de información mutua condicionada. Los estimadores propuestos permiten por una parte adaptar los algoritmos que, estando basados en BNs o CGNs, emplean la cantidad de informacion mutua; y por otra, diseñar medidas filter [11] para la selección de variables. El trabajo finaliza con la Sección 5 en la que se citan las principales aportaciones del trabajo.

2. Estimadores de densidad basados en kernels Gaussianos esféricos

El estimador d-dimensional basado en kernels en su forma más general es

$$f(\boldsymbol{x};\boldsymbol{H}) = n^{-1}\sum_{i=1}^{n} K_{\boldsymbol{H}}(\boldsymbol{x}-\boldsymbol{x^{(i)}}) \qquad (2)$$

donde $\boldsymbol{H}$ es la matriz de ancho de banda (*BM*, *bandwidth matrix*) y n el número de casos con los que se construye el estimador,

$$K_{\boldsymbol{H}}(\boldsymbol{x}) = |\boldsymbol{H}|^{-1/2}K(\boldsymbol{H}^{-1/2}\boldsymbol{x}) \qquad (3)$$

asumiendo que K es una función de densidad d-variada. Un estimador kernel viene dado por dos parámetros:

1. El kernel K seleccionado
2. La matriz de ancho de banda $\boldsymbol{H}$

El kernel seleccionado para el paradigma CFB es la función de densidad normal d-dimensional con matriz de covarianzas $S = I$

$$\begin{aligned} K(\boldsymbol{x}) &= (2\pi)^{-d/2}exp(-1/2\boldsymbol{x}^T\boldsymbol{x}) \\ &\sim N(\boldsymbol{0},\boldsymbol{I}) \end{aligned} \qquad (4)$$

en cuyo caso $K_{\boldsymbol{H}}(\boldsymbol{x}-\boldsymbol{x^{(i)}})$ equivale a la función de densidad $N(\boldsymbol{x^{(i)}},\boldsymbol{H})$. $\boldsymbol{H}$ es una matriz simétrica $d \times d$, por lo que en general posee $\frac{d(d+1)}{2}$ parámetros diferentes. Este número de parámetros puede ser excesivamente elevado, incluso para bajas dimensiones, lo que sugiere restringir $\boldsymbol{H}$ para que sea más simple (matriz escalar, diagonal,...).

Antes de continuar con el problema de seleccionar el $\boldsymbol{H}$, presentaremos dos transformaciones clásicas sobre el espacio de variables, que serán útiles en la siguiente subsección:

1. Escalado (*scaling*)

$$\boldsymbol{X}^* = diag(s_1,...,s_d)\boldsymbol{X} \qquad (5)$$

donde $diag(\boldsymbol{s})$ es una matriz diagonal, y s_i un estimador de la desviación típica. Equivale a transformar los datos para que todas las variables tengan una desviación típica unitaria.

2. Transformación esférica (*sphering*)

$$\boldsymbol{X}^* = S^{-1/2}\boldsymbol{X} \quad (6)$$

donde S es una matriz de dispersión. Nosotros emplearemos el estimador muestral de la matriz de covarianzas de $\boldsymbol{X}$. Esto a transformar linealmente los datos para que tengan una matriz de covarianzas $S^* = I$, es decir, elimina las diferencias de escala y la correlación entre variables [18].

2.1. Seleccionando $\boldsymbol{H}$

La selección de la (*BM*) óptima $\boldsymbol{H}$ es crucial para conseguir una buena estimación de una función de densidad, y es más determinante incluso que la elección del kernel $K(\cdot)$ que se emplee [4]. $\boldsymbol{H}$ establece el grado de suavizado de la estimación de la función de densidad. En el caso univariado el estimador depende de un único parámetro h, que hace las veces de la *BM* $\boldsymbol{H}$ en el caso d-variado. De forma intuitiva, partiendo de valores de h cercanos a cero obtenemos una estimación muy ruidosa, con muchas fluctuaciones. Conforme h va aumentando comienzan a desaparecer los ruidos y comienza a aproximarse a la densidad real, hasta que se alcanza el óptimo. Conforme h sigue aumentando, y alejándose del óptimo, comienzan a perderse detalles de la función debido al *sobre-suavizado* (*oversmoothed*). Conforme h se aproxima a ∞ la función de densidad se vuelve más y más plana y el estimador comienza a parecerse al *vecino más cercano* (*nearest neighbor*).

El número de parámetros a estimar en una *BM* completa es $\mathcal{O}(d^2)$. Por tanto, el problema se vuelve inmanejable rápidamente conforme crece d lo que sugiere que es necesario restringir $\boldsymbol{H}$. Las posibilidades que se consideran habitualmente son tres [18]:

1. $\boldsymbol{H} = h^2 I$. Al mantener constante el parámetro de suavizado h para todas las variables suavizamos la estimación de la densidad de todas por igual. Se recomienda escalar previamente las variables.

2. $\boldsymbol{H} = diag(h_1^2, ..., h_d^2) = h^2 diag(s_1^2, ..., s_d^2)$. h_i es el parámetro de suavizado de la variable X_i y s_i es su constante de escalado (desviación típica,...). Es equivalente a escalar primero las variables y emplear la primera aproximación a la BM óptima.

3. $\boldsymbol{H} = h^2 S$. Es equivalente a transformar esféricamente los datos y suavizarlos con h mediante la primera opción. Intuitivamente, se trata de utilizar un kernel en la estimación con la misma forma que la densidad real.

Por tanto las tres opciones se pueden convertir en la primera empleando correctamente la transformaciones esféricas (Ecuación 6) y/o el escalado (Ecuación 5).

El estimador que será empleado por las CFN equivale a transformar esféricamente los datos, y emplear el kernel de la Ecuación 4 con una BM $\boldsymbol{H} = h^2 I$. Este estimador fue propuesto por Fukunaga [9] y viene dado por la expresión [17]

$$\hat{f}(\boldsymbol{x}) = \frac{|S|^{-1/2}}{nh^d} \sum_{i=1}^{n} K(hS^{1/2}(\boldsymbol{x} - \boldsymbol{x}^{(i)})) \quad (7)$$

A pesar de que la intuición nos dice que estamos ante una buena elección de $\boldsymbol{H}$, escoger siempre la transformación esférica está desaconsejada por Wand y Jones [19] para la estimación de densidades. A pesar de ello tenemos la creencia de que se obtendrán mejores resultados empleando el estimador de Fukunaga ya que Duong y Hazelton [7] emplean la transformación esférica y obtienen buenos resultados en la mayoría de sus estimaciones.

El principal problema, en la mayoría de situaciones prácticas, es que la función de densidad es desconocida por lo que no es posible optimizar ningún criterio de pérdida (una especie de distancia entre la estimación y la función de densidad real). En la literatura se ha tratado el problema intentando minimizar criterios relacionados con el *error cuadrático*, ya que el análisis matemático del error cuadrático

parece ser el más sencillo. El criterio más empleado es el error cuadrático integrado medio (*MISE, mean integrated squared error*), que se define de la siguiente manera:

$$MISE\hat{f}(\boldsymbol{X}, h) = \int E\{\hat{f}(\boldsymbol{x}, h) - f(\boldsymbol{x}, h)\}^2 d\boldsymbol{x}$$

En este trabajo proponemos dos opciones para seleccionar la *BM* $\boldsymbol{H}$ basadas en el MISE asintótico bajo la suposición de que $\boldsymbol{H} = h^2 I$ (*AMISE, asymptotic mean integrated squared error*), los cuales se complementan a la perfección:

1. Regla normal (*normal rule*) de Fukunaga [9].
2. Regla DPI (*Direct plug-in rule*) de Duong y Hazelton [7].

Para poder derivar la expresión del AMISE de un estimador (empleando la expansión de las series de Taylor) es necesario asumir que todas las derivadas parciales segundas son continuas por partes e integrables, y que el kernel satisface las siguientes condiciones: $(i,j)_1^d \lim_{n\to\infty}(h_{i,j}) = 0$, $\lim_{n\to\infty}(n^{-1}|\boldsymbol{H}|^{-1/2}) = 0$, $\int K(\boldsymbol{x})d\boldsymbol{x} = 1$, $\int \boldsymbol{x}K(\boldsymbol{x})d\boldsymbol{x} = 0$ y $\int \boldsymbol{x}\boldsymbol{x}^T K(\boldsymbol{x})d\boldsymbol{x} = I$ [21]. El kernel que emplea el estimador de Fukunaga [9] (Ecuación 7) cumple todas estas restricciones.

La *regla normal* selecciona la BM $\boldsymbol{H}$ óptima desde el punto de vista del AMISE y bajo la suposición de que la densidad real de las variables sigue una normal multidimensional con la matriz de covarianzas identidad I. Bajo esta suposición, encontramos una BM diagonal cuyos elementos se calculan mediante la expresión:

$$h_{i,j} = \begin{cases} (\frac{4}{(d+2)})^{\frac{1}{d+4}} \sigma_i n^{-\frac{1}{d+4}} & \text{si } i = j \\ 0 & \text{en otro caso} \end{cases}$$

donde σ_i^2 es la varianza de la variable i-ésima (elemento (i,i) de S). Empleando esta regla obtenemos $\boldsymbol{H} = h^2 I$ lo que equivale a realizar el mismo escalado sobre todas las variables. Para obtener una matriz $\boldsymbol{H}$ completa que tenga en cuenta las relaciones de dependencia que existen entre las variables empleamos la *transformación esférica* de la Ecuación 6 sobre los datos, calculamos $\boldsymbol{H}^* = (h^*)^2 I$ y aplicamos la transformación inversa $\boldsymbol{H} = S^{1/2}\boldsymbol{H}^* S^{1/2}$.

Las reglas DPI [22] explotan la relación que existe entre el error cuadrático medio (*MSE, mean squared error*), el *sesgo*[2] y la varianza[3]. La expresión del sesgo de un estimador está en función de la densidad a estimar, que usualmente se aproxima empleando las expansiones de las series de Taylor. La idea que subyace a las reglas DPI es obtener un estimador del sesgo sustituyendo un estimador de la densidad real en una expresión aproximada del sesgo. A partir del estimador del sesgo y del estimador de la varianza se puede obtener una expresión cerrada para la estimación del MISE. La BM $\boldsymbol{H} = h^2 I$ óptima se obtiene minimizando la expresión del estimador del MISE con respecto a h. Muchas son las reglas DPI desarrolladas, entre las cuales la regla de Wand and Jones [20] es una de las más conocidas.

La regla DPI de Duong y Hazelton [7] que empleamos tiene la ventaja con respecto de la regla DPI de Wand and Jones [20] de que es más estable, ya que siempre encuentra BMs finitas[4]. Además, la aproximación de Duong y Hazelton requiere menos cómputo, siendo independiente de la dimensión de la densidad que se quiere estimar. Estas ventajas se consiguen gracias a que Duong y Hazelton siguen un criterio de optimización diferente. Wand y Jones [20] intentan que su regla esté calibrada para optimizar la estimación de la matriz de funciones $\boldsymbol{\Psi}$ de dimensiones $(\frac{(d+1)d}{2} \times \frac{(d+1)d}{2})$ elemento por elemento, siendo cada elemento de la forma $\psi_{\boldsymbol{r}} = \int_{\Re^d} f^{(\boldsymbol{r})}(\boldsymbol{x}) f(\boldsymbol{x}) d\boldsymbol{x}$. Duong y Hazelton [7] optimizan la estimación de $\boldsymbol{\Psi}$ empleando un único parámetro, y para ello introducen el concepto de minimizar la suma de AMSEs (*SAMSE*) de los estimadores de las funciones $\psi_{\boldsymbol{r}}$. Duong y Hazelton recomiendan computar su regla DPI en dos pasos, $l = 2$ [7].

Ambas opciones se complementan de la siguiente manera. La regla normal es eficiente

[2] El sesgo (*bias*) se define como: $bias(\hat{f}(\boldsymbol{x})) = E\hat{f}(\boldsymbol{x}) - f(\boldsymbol{x})$

[3] $MSE(\hat{f}(\boldsymbol{x})) = Var(\hat{f}(\boldsymbol{x})) + bias(\hat{f}(\boldsymbol{x}))^2$

[4] Con elementos distintos de ∞ y $-\infty$

ya que tan solo requiere computar una expresión cerrada cuyo orden de complejidad computacional es $\mathcal{O}(d)$. La regla normal tiene la desventaja de que generalmente tiende a sobresuavizar la estimación, por lo que es posible que se pierdan características importantes de la densidad real. La regla DPI de Duong y Hazelton aborda el problema empleando suposiciones más débiles mediante las cuales se obtiene un estimador que se ajusta mejor a la densidad real de las variables. Esta regla DPI posee un coste computacional $\mathcal{O}(n^2 l)$ (independiente de la dimensionalidad d).

3. Redes flexibles condicionadas

Una CFN es un PGM, es decir, un grafo compuesto por un conjunto de arcos que representan relaciones de (in)dependencia condicional, nodos representando las variables del problema, y un conjunto de parámetros para modelar las relaciones de (in)dependencia del grafo. Posee la misma restricción estructural que las CGN: un nodo discreto no puede tener padres continuos. La diferencia principal con respecto de las BNs y CGNs se encuentra en la modelización que se hace del grafo, es decir, de las densidades que siguen las variables aleatorias condicionadas a sus padres. Siendo $\boldsymbol{X}$ e $\boldsymbol{Y}$ variables aleatorias multidimensionales continuas y $\boldsymbol{Z}$ una variable multidimensional discreta, la estimación de la función de densidad $f(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{z})$ viene dada por la expresión:

$$\hat{f}(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{z}) = \hat{f}_{\boldsymbol{z}}(\boldsymbol{x}|\boldsymbol{y}) = \frac{\hat{f}_{\boldsymbol{z}}(\boldsymbol{x},\boldsymbol{y})}{\hat{f}_{\boldsymbol{z}}(\boldsymbol{y})} = C^{\boldsymbol{z}}_{\boldsymbol{X}|\boldsymbol{Y}} \cdot \frac{\sum_{i=1}^{n} K((\boldsymbol{H}^{\boldsymbol{z}}_{(\boldsymbol{X},\boldsymbol{Y})})^{-1/2}((\boldsymbol{x},\boldsymbol{y}) - (\boldsymbol{x}_i,\boldsymbol{y}_i)))}{\sum_{i=1}^{n} K((\boldsymbol{H}^{\boldsymbol{z}}_{\boldsymbol{Y}})^{-1/2}(\boldsymbol{y} - \boldsymbol{y}_i))} \quad (8)$$

donde $f(\cdot)$ sigue una densidad de Fukunaga (Ecuación 7) y $f(\cdot)_{\boldsymbol{z}}$ es la función de densidad computada con la partición de los casos en los que $\boldsymbol{Z} = \boldsymbol{z}$. $C^{\boldsymbol{z}}_{\boldsymbol{X}|\boldsymbol{Y}}$ es un coeficiente dependiente del valor $\boldsymbol{z}$ de la instancia (independiente de los valores $(\boldsymbol{x},\boldsymbol{y})$) y dependiente de las variables $\boldsymbol{X}$ e $\boldsymbol{Y}$. Este coeficiente viene dado por la expresión:

$$C^{\boldsymbol{z}}_{\boldsymbol{X}|\boldsymbol{Y}} = \frac{(h^{\boldsymbol{z}}_{\boldsymbol{Y}})^{d_Y}}{(2\pi)^{d_X/2}(h^{\boldsymbol{z}}_{(\boldsymbol{X},\boldsymbol{Y})})^{d_{\boldsymbol{X}}+d_{\boldsymbol{Y}}}} \cdot \{\frac{|S^{\boldsymbol{z}}_{\boldsymbol{Y}}|}{|S^{\boldsymbol{z}}_{(\boldsymbol{X},\boldsymbol{Y})}|}\}^{1/2} \quad (9)$$

Tal y como se expuso en la Sección 2, las BMs que se emplean en las CFNs están restringidas a $\boldsymbol{H}^{\boldsymbol{z}}_{\boldsymbol{U}} = (h^{\boldsymbol{z}}_{\boldsymbol{U}})^2 S^{\boldsymbol{z}}_{\boldsymbol{U}}$. Por tanto, la función de densidad de la Ecuación 8 depende de $\mathcal{O}(|\boldsymbol{Z}|)$ parámetros de suavizado: $\forall \boldsymbol{z} = 1,...,|\boldsymbol{Z}|; h^{\boldsymbol{z}}_{(\boldsymbol{x},\boldsymbol{y})}$ para ajustar $f_{\boldsymbol{z}}(\boldsymbol{x},\boldsymbol{y})$ y $h^{\boldsymbol{z}}_{\boldsymbol{y}}$ para ajustar $f_{\boldsymbol{z}}(\boldsymbol{y})$.

La elección de los parámetros $h^{\boldsymbol{z}}_{(\boldsymbol{X},\boldsymbol{Y})}$ y $h^{\boldsymbol{z}}_{\boldsymbol{Y}}$ se complica puesto que tratamos de minimizar el error del cociente de $\hat{f}_{\boldsymbol{z}}(\boldsymbol{x},\boldsymbol{y})$ entre $\hat{f}_{\boldsymbol{z}}(\boldsymbol{y})$. Los valores de $h^{\boldsymbol{z}}_{(\boldsymbol{X},\boldsymbol{Y})}$ y de $h^{\boldsymbol{z}}_{\boldsymbol{Y}}$ que minimizan $MISE(\hat{f}_{\boldsymbol{z}}(\boldsymbol{x},\boldsymbol{y}))$ y $MISE(\hat{f}_{\boldsymbol{z}}(\boldsymbol{y}))$ parecen obtener un buen valor de $MISE(\hat{f}_{\boldsymbol{z}}(\boldsymbol{x}|\boldsymbol{y}))$ ya que

$$MISE(\hat{f}_{\boldsymbol{z}}(\boldsymbol{x}|\boldsymbol{y})) = \int_{\Re^d} [\frac{f_{\boldsymbol{z}}(\boldsymbol{x},\boldsymbol{y})\hat{f}_{\boldsymbol{z}}(\boldsymbol{y})}{f_{\boldsymbol{z}}(\boldsymbol{y})\hat{f}_{\boldsymbol{z}}(\boldsymbol{y})} - \frac{\hat{f}_{\boldsymbol{z}}(\boldsymbol{x},\boldsymbol{y})f_{\boldsymbol{z}}(\boldsymbol{y})}{f_{\boldsymbol{z}}(\boldsymbol{y})\hat{f}_{\boldsymbol{z}}(\boldsymbol{y})}]^2 d\boldsymbol{x}d\boldsymbol{y} \quad (10)$$

Teniendo en cuenta que si obtenemos el $h_{\boldsymbol{U}}$ que optimiza $MISE(h_{\boldsymbol{U}})$ tenemos que $\hat{f}_{h_{\boldsymbol{U}}}(\boldsymbol{u}) \simeq f(\boldsymbol{u})$ y podemos deducir que $f_{\boldsymbol{z}}(\boldsymbol{x},\boldsymbol{y})\hat{f}_{\boldsymbol{z}}(\boldsymbol{y}|\boldsymbol{z}) \simeq \hat{f}_{\boldsymbol{z}}(\boldsymbol{x},\boldsymbol{y})f_{\boldsymbol{z}}(\boldsymbol{y})$. Por tanto consideramos que la elección de los parámetros obtendrá un valor para $MISE(\hat{f}_{\boldsymbol{z}}(\boldsymbol{x}|\boldsymbol{y}))$ al menos aceptable. En la implementación del FB [10] se utilizan los parámetros de suavizado $\forall i = 1,...,d; h_i = \sigma_i$ y el clasificador se comporta bien. Por lo tanto, cabe esperar que se obtengan mejores resultados, seleccionando los parámetros de suavizado de tal manera que minimicen $MISE(\hat{f}_{\boldsymbol{z}}(\boldsymbol{x},\boldsymbol{y}))$ y $MISE(\hat{f}_{\boldsymbol{z}}(\boldsymbol{y}))$.

Modelar un clasificador basado en PGM equivale a modelar la factorización del manto de Markov (*Markov blanket*) de la variable clase. En la factorización del manto de Markov de C únicamente intervienen un máximo de d factores, por lo que es necesario estimar un máximo de $\mathcal{O}(d|C|)$ parámetros de suavizado. Hemos decidido emplear la regla DPI de Duong y Hazelton [7] en dos pasos ($l = 2$)

para obtener los parámetros de suavizado, por lo que el costo computacional para modelar una estructura arbitraria basada en CFNs es como máximo $\mathcal{O}(n^2 d|C|)$.

4. Adaptando el algoritmo TAN de Friedman y col. (1997)

En esta Sección presentamos y adaptamos el algoritmo TAN de Friedman y col. [8] que induce redes con *estructura naive Bayes aumentada a árbol*. Este tipo de estructuras se obtienen construyendo primero una estructura de árbol entre las predictoras para posteriormente unir la variable clase con cada una de las variables predictoras.

El algoritmo TAN de inducción de clasificadores consiste básicamente en una adaptación del algoritmo de Chow y Liu [1], en la cual se emplea la cantidad de información mutua condicionada a la clase en lugar de la cantidad de información mutua [2].

A continuación proponemos los estimadores para la cantidad de información mutua entre dos variables continuas y para la cantidad de información mutua condicionada a una variable discreta. Ambos estimadores están en su forma más general (para variables multidimensionales). El objetivo es presentar un estimador que permita cuantificar la cantidad de información mutua entre dos variables continuas cualesquiera. Esta formulación será convenientemente adaptada a las características concretas del algoritmo TAN, pero de la misma manera se podría adecuar a las necesidades de otros algoritmos de inducción de clasificadores, incluidos aquellos en los que intervienen variables multidimensionales (introducidas para el algoritmo *semi naive Bayes* [13]).

4.1. Cantidad de información mutua y cantidad de información mutua condicionada

La cantidad de información mutua entre dos variables continuas multidimensionales $\boldsymbol{X}$ e $\boldsymbol{Y}$, cuya densidad viene dada por la estimación basada en kernels, está definida como [2]

$$I(\boldsymbol{X},\boldsymbol{Y}) = \int \hat{f}(\boldsymbol{x},\boldsymbol{y}) log \frac{\hat{f}(\boldsymbol{x},\boldsymbol{y})}{\hat{f}(\boldsymbol{x})\hat{f}(\boldsymbol{y})} d\boldsymbol{x} d\boldsymbol{y}$$
$$= E(log \frac{\hat{f}(\boldsymbol{x},\boldsymbol{y})}{\hat{f}(\boldsymbol{x})\hat{f}(\boldsymbol{y})}) \quad (11)$$

Dada la definición de la Ecuación 11, proponemos el siguiente estimador muestral

$$\hat{I}(\boldsymbol{X},\boldsymbol{Y}) = \frac{1}{n}\sum_{i=1}^{n} log \frac{\hat{f}(\boldsymbol{x}^{(i)},\boldsymbol{y}^{(i)})}{\hat{f}(\boldsymbol{x}^{(i)})\hat{f}(\boldsymbol{y}^{(i)})} \quad (12)$$

La cantidad de información mutua entre dos variables continuas condicionada a una variable multinomial multidimensional $\boldsymbol{Z}$ viene dada por

$$I(\boldsymbol{X},\boldsymbol{Y}|\boldsymbol{Z}) = \sum_{\boldsymbol{z}=1}^{|\boldsymbol{Z}|} p(\boldsymbol{z}) I_{\boldsymbol{z}}(\boldsymbol{X},\boldsymbol{Y}) \quad (13)$$

Siendo $I_{\boldsymbol{z}}(\boldsymbol{X},\boldsymbol{Y}) = E(log \frac{\hat{f}_{\boldsymbol{z}}(\boldsymbol{x},\boldsymbol{y})}{\hat{f}_{\boldsymbol{z}}(\boldsymbol{x})\hat{f}_{\boldsymbol{z}}(\boldsymbol{y})})$ donde $\hat{f}_{\boldsymbol{z}}(\cdot)$ es el estimador basado en kernels empleando únicamente los casos de la partición con $\boldsymbol{Z} = \boldsymbol{z}$. A partir de las Ecuaciones 12 y 13 proponemos el estimador

$$\hat{I}(\boldsymbol{X},\boldsymbol{Y}|\boldsymbol{Z}) = \sum_{\boldsymbol{z}=1}^{|\boldsymbol{Z}|} p(\boldsymbol{z}) \hat{I}_{\boldsymbol{z}}(\boldsymbol{X},\boldsymbol{Y}) \quad (14)$$
$$= \sum_{\boldsymbol{z}=1}^{|\boldsymbol{Z}|} p(\boldsymbol{z}) \frac{1}{n_{\boldsymbol{z}}} \sum_{i=\boldsymbol{z}:1}^{\boldsymbol{z}:n} log \frac{\hat{f}_{\boldsymbol{z}}(\boldsymbol{x}^{(i)},\boldsymbol{y}^{(i)})}{\hat{f}_{\boldsymbol{z}}(\boldsymbol{x}^{(i)})\hat{f}_{\boldsymbol{z}}(\boldsymbol{y}^{(i)})}$$

donde el superíndice $(\boldsymbol{z} : j)$ hace referencia al j-ésimo caso de la partición en la que los casos toman el valor $\boldsymbol{Z} = \boldsymbol{z}$ y $n_{\boldsymbol{z}}$ es el número de casos en la misma partición.

4.2. Adaptación

En esta subsección presentamos el algoritmo TAF, adaptación del TAN al paradigma CFN.

El pseudocódigo del TAN de Friedman y col. (1997) se muestra en el Algoritmo 1. Para implementar el TAF es necesario emplear el estimador de la Ecuación 14 con predictoras continuas unidimensionales $\boldsymbol{X} = X_i$ y $\boldsymbol{Y} = X_j$, condicionadas a la clase $\boldsymbol{Z} = C$. La expresión

para la información mutua viene dada por la Ecuación 15.

$$\hat{I}\left(X_j, X_k|C\right) = \frac{1}{n}\sum_{c=1}^{|C|} p(c) \sum_{i=c:1}^{c:n} log\frac{\hat{f}_c(x_j^{(i)}, x_k^{(i)})}{\hat{f}_c(x_j^{(i)})\hat{f}_c(x_k^{(i)})} \quad (15)$$

1. Calcular $I(X_i, X_j \mid C)$ con $i < j, i, j = 1, \ldots, d$
2. Construir un grafo no dirigido completo cuyos nodos corresponden a las variables predictoras: $X_1, \ldots, X_d$. Asignar a cada arista conectando las variables X_i y X_j un peso dado por $I(X_i, X_j \mid C)$
3. A partir del grafo completo anterior y siguiendo el algoritmo de Kruskall construir un árbol expandido de máximo peso
4. Transformar el árbol no dirigido resultante en uno dirigido, escogiendo una variable como raíz, para a continuación direccionar el resto de aristas
5. Construir un modelo *TAN* añadiendo un nodo clase C y posteriormente un arco desde C a cada variable

Algoritmo 1: *TAN* (Friedman y col. 1997).

Para conseguir que el algoritmo CFN sea computacionalmente viable es vital seleccionar la estimación de los parámetros de suavizado adecuada.

A pesar de que en la Ecuación 15 no se especifique explícitamente, $\hat{f}_c(x_j^{(i)}, x_k^{(i)})$, $\hat{f}_c(x_j^{(i)})$ y $\hat{f}_c(x_k^{(i)})$ dependen de la elección de la MB. Para lograr que TAF sea viable computacionalmente es vital realizar una estimación eficiente de los parámetros de suavizado para computar la cantidad de información mutua de la Ecuación 15, ya que es necesario computar $\mathcal{O}(d^2)$ cantidades diferentes. Por lo tanto, en el primer paso del algoritmo, se ha decidido emplear la regla normal debido a su simplicidad computacional. Una vez obtenido el modelo se crea la factorización asociada y siguiendo la pauta introducida en la Sección 3 se reestiman $\mathcal{O}(d)$ parámetros de suavizado mediante la regla DPI de Duong y Hazelton [7].

Siguiendo estas pautas para la selección de los parámetros de suavizado, la complejidad computacional del algoritmo TAF es $\mathcal{O}(|C|n^2d + d^2\prod_{c=1}^{|C|} n_c^2)$. Para el algoritmo FB (seleccionando $\boldsymbol{H}$ empleando la regla DPI [7] con $l = 2$) el costo computacional en el aprendizaje es $\mathcal{O}(n^2d|C|)$, mientras que en el caso del algoritmo TAN es tan solo $\mathcal{O}(d^2|C| + n)$. La clasificación de una instancia empleando el TAF requiere $\mathcal{O}(dn|C|)$ operaciones (el mismo orden que el algoritmo FB), mientras que el TAN requiere únicamente de $\mathcal{O}(d|C|)$ operaciones.

5. Principales aportaciones

Este trabajo puede ser considerado como la presentación del nuevo paradigma *red flexible condicionada*, no pretende ser un estudio en profundidad del mismo. La CFN, a grandes rasgos, es un PGM que asume que la densidad de las variables continuas está basada en el estimador de Fukunaga [17] (Ecuación 7). Este paradigma puede ser considerado como una extensión de las BNs y RCGs desde el punto de vista de la flexibilidad de los estimadores de densidad que emplea.

A modo de ejemplo se incluye la adaptación del algoritmo TAN basado en las BNs de Friedman y col. (1997), y puede ser considerado como una extensión del clasificador *FB* de John y Langley (1995) desde el punto de vista de las relaciones de (in)dependencia condicional permitidas.

Además, como elementos indispensables para implementar los algoritmos basados en información mutua, se incluyen los estimadores para la cantidad de información mutua y la cantidad de información mutua condicionada entre dos variables multidimensionales con una función de densidad basada en kernels.

6. Agradecimientos

Este trabajo ha sido posible gracias a la beca del Gobierno Vasco para la Formación de

Investigadores 2004-05, a la Universidad del País Vasco con la beca 9/UPV 00140.226-15334/2003 y a los proyectos del Gobierno Vasco SAITEK S-PE04UN25, ETORTEK-GENMODIS y ETORTEK-BIOLAN.

Referencias

[1] C. Chow y C. Liu. Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, 14:462–467, 1968.

[2] T. M. Cover y J. A. Thomas. *Elements of Information Theory*. John Wiley and Sons, 1991.

[3] M. DeGroot. *Optimal Statistical Decisions*. McGraw-Hill, New York, 1970.

[4] A. Delaigle y I. Gijbels. Comparison of data-driven bandwidth selection procedures in deconvolution kernel density estimation. *Computational Statistics and Data Analysis*, pages 1–20, 2002.

[5] P. Domingos y M. Pazzani. On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning*, 29:103–130, 1997.

[6] R. Duda y P. Hart. *Pattern Classification and Scene Analysis*. John Wiley and Sons, 1973.

[7] T. Duong y M.l. Hazelton. Plug-in bandwidth matrix for bivariate kernel density estimation. *Nonparametric Statistics*, 15(1):17–30, 2003.

[8] N. Friedman, D. Geiger, y M. Goldszmidt. Bayesian network classifiers. *Machine Learning*, 29:131–163, 1997.

[9] K. Fukunaga. *Statistical Pattern Recognition*. Academic Press, Inc, 1972.

[10] G. John y P. Langley. Estimating continuous distributions in Bayesian classifiers. In *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence*, pages 338–345, 1995.

[11] H. Liu y H. Motoda. *Feature Selection for Knowledge Discovery and Data Mining*. Kluwer Academic Publishers, 1998.

[12] S. L. Lauritzen y D. J. Spiegelhalter. Mixed interaction models. Technical report r 84-8, Institute for Electronic Systems, Aalborg University, 1984.

[13] M. Pazzani. Searching for dependencies in Bayesian classifiers. In *Learning from Data: Artificial Intelligence and Statistics V*, pages 239–248, 1997.

[14] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers, 1988.

[15] A. Pérez, P. Larrañaga, e I. Inza. Supervised classification with conditional Gaussian networks: increasing the structure complexity from naive Bayes. *Submitted*, 2005.

[16] M. Rosenblatt. Remarks on some nonparametric estimates of a density function. *Annals of Mathematical Statistics*, 27:832–837, 1956.

[17] B.W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, 1986.

[18] J.S Simonoff. *Smoothing Methods in Statistics*. Springer, 1996.

[19] M.P. Wand y M.C. Jones. Comparison of smoothing parametrizations in bivariate kernel density estimation. *Journal of the American Statistical Association*, 88(422):520–528, 1993.

[20] M.P. Wand y M.C. Jones. Multivariate plug-in bandwidth selection. *Computational Statistics*, 9:97–116, 1994.

[21] M.P. Wand y M.C. Jones. *Kernel Smoothing*. Monographs on Statistics and Applied Probability. Cahpman and Hall, 1995.

[22] M. Woodroofe. On choosing a delta sequence. *Annals of Mathematical Statistics*, 41:1665–1671, 1970.