Aprendizaje discriminativo de clasificadores Bayesianos

Guzmán Santafé, Jose A. Lozano, Pedro Larrañaga
Grupo de Sistemas Inteligentes
Dept. de Ciencias de la Computación e Inteligencia Artificial
Universidad del País Vasco
{guzman, lozano, ccplamup}@si.ehu.es

Resumen

El aprendizaje de modelos de clasificación Bayesianos es, generalmente, un proceso generativo. Es decir, en el aprendizaje de dichos modelos se busca maximizar la verosimilitud del conjunto de datos dado el modelo aprendido. Sin embargo, existe otra aproximación, el aprendizaje discriminativo, en la que se busca maximizar la verosimilitud condicional. Ésta parece, en principio, una aproximación más natural para propósitos de clasificación. No obstante, no siempre los modelos aprendidos mediante un método discriminativo son mejores que los aprendidos con métodos generativos. Recientemente han aparecido en la bibliografía varios métodos para el aprendizaje discriminativo de clasificadores Bayesianos. En el presente artículo, se presenta un nuevo método para el aprendizaje discriminativo, tanto de la estructura como de los parámetros, de clasificares Bayesianos. Esta aproximación se basada en la adaptación del algoritmo TM a modelos de clasificación Bayesianos. Por otra parte, en el presente artículo, también se presenta una evaluación experimental del método propuesto aplicado a diferentes bases de datos estándares para clasificación supervisada.

1. Introducción

La clasificación supervisada es una parte del aprendizaje automático con gran número de aplicaciones en distintos campos como son el reconocimiento de patrones o el diagnóstico médico. Generalmente, en la clasificación supervisada se asume la existencia de dos tipos de variables: las variables predictoras, $\boldsymbol{X} = (X_1, \ldots, X_n)$, y la variable clase, C. Mediante los clasificadores supervisados se trata de aprender las relaciones entre las variables predictoras y la clase, de forma que se pueda asignar un valor de C a un nuevo caso, $\boldsymbol{x} = (x_1, \ldots, x_n)$, en el que el valor de la clase es desconocido.

Uno de los paradigmas ampliamente usados para clasificación supervisada son las redes Bayesianas [7]. Éstas son modelos gráficos probabilísticos [13] que permiten modelar de una forma simple y precisa la distribución de probabilidad subyacente a un conjunto de datos. Además, la representación gráfica de las relaciones de dependencia entre las variables presentes en el conjunto de datos se realiza mediante un grafo acíclico dirigido lo cual facilita la comprensión e interpretabilidad del modelo de clasificación. Generalmente, las redes Bayesianas, son consideradas clasificadores generativos, ya que en el aprendizaje de dichos clasificadores se busca maximizar la verosimilitud conjunta, es decir, se busca el modelo que mejor representa al conjunto de datos. Por el contrario, existe otra tendencia que es el aprendizaje discriminativo, en el cual se busca maximizar la verosimilitud condicional.

El aprendizaje discriminativo parece una aproximación más natural al aprendizaje de modelos de clasificación, ya que en el aprendizaje está guiado precisamente por la probabilidad condicional, $p(C|X_1, \ldots, X_n)$, y es este valor de probabilidad el que, una vez aprendido el modelo, se utiliza para clasificar nuevas

Actas del III Taller Nacional de Minería de Datos y Aprendizaje, TAMIDA2005, pp.115-123
ISBN: 84-9732-449-8

instancias en las que el valor de la variable clase es desconocido. Desafortunadamente, mientras que el cálculo de la verosimilitud conjunta puede ser obtenido mediante una fórmula cerrada, lo cual permite un cómputo eficiente, en el cálculo de la verosimilitud condicional ésto no es posible. Por tanto el aprendizaje discriminativo de modelos de clasificación Bayesianos es más costoso que la correspondiente aproximación generativa.

Recientemente se han propuesto diversas técnicas para afrontar el problema del aprendizaje discriminativo de clasificadores Bayesianos una vez la estructura del modelo ha sido fijada. Por una parte en [8, 14] se proponen métodos genéricos de optimización numérica para la obtención de los parámetros del modelo que maximizan la verosimilitud condicional. Por otra, en [17] el método de optimización para el aprendizaje discriminativo de los parámetros del modelo se propone desde un punto de vista estadístico, basado en estadísticos suficientes del conjunto de datos.

Como se apuntaba anteriormente, el aprendizaje discriminativo de los parámetros en modelos de clasificación Bayesianos es un tema recientemente tratado en la literatura. Por el contrario, el aprendizaje discriminativo del conjunto de estructura y parámetros del clasificador apenas se ha abordado. Hasta donde nosotros conocemos, sólo [9] propone un método en el que se combina un aprendizaje generativo de los parámetros del modelo con una selección de la estructura en base a la verosimilitud condicional. Por otra parte, también propone utilizar el método de [8] para el aprendizaje discriminativo de los parámetros y así conseguir un aprendizaje discriminativo tanto de la estructura como de los parámetros. Sin embargo, esta solución resulta muy costosa e incluso computacionalmente inviable para bases de datos relativamente grandes.

En el presente artículo se propone una extensión del aprendizaje discriminativo de clasificadores Bayesianos propuesto en [17] para abordar el aprendizaje estructural de dichos clasificadores desde un punto de vista discriminativo. De esta forma, se permite una optimización completa, tanto de estructura como de parámetros, desde una aproximación discriminativa.

El resto del artículo está organizado de la siguiente forma: la Sección 2 hace una pequeña introducción a los modelos de clasificación Bayesianos clásicos, generalmente considerados modelos generativos. En la Sección 3 se presenta la idea del algoritmo TM como método discriminativo de aprendizaje de los parámetros para clasificadores Bayesianos y también el uso de este algoritmo para el aprendizaje discriminativo de la estructura del clasificador. Método, éste último, que hemos bautizado con el nombre de TM estructural. A continuación, en la Sección 4 se presenta una evaluación experimental del algoritmo TM estructural utilizando diferentes bases de datos estándar. Por último, en la Sección 5, exponemos las conclusiones obtenidas del presente artículo así como las líneas futuras de investigación.

2. Clasificadores Bayesianos generativos

Los modelos de clasificación Bayesianos surgen del uso de las redes Bayesianas para propósitos clasificatorios. Dependiendo de las restricciones que se le ponga a la red Bayesiana se han propuesto diferentes clasificadores en la bibliografía. Por ejemplo, el modelo naive Bayes [12] es el clasificador Bayesiano más sencillo. Éste se construye bajo la suposición de que todas las variables predictoras son condicionalmente independientes dado el valor de la variable clase. El modelo naive Bayes tiene una estructura fija que no depende de los datos, por tanto no podemos decir que el aprendizaje estructural de este modelo siga un criterio generativo o discriminativo. Sin embargo, el aprendizaje habitualmente utilizado para obtener los parámetros del modelo naive Bayes, así como el de la mayoría de clasificadores Bayesianos en general, está basado en las estimaciones *máximo verosímil* (ML) o *máximo a posteriori* (MAP) [2, 10] del conjunto de parámetros. Éstas sí son aproximaciones generativas al aprendizaje de los parámetros del clasificador ya que con ellas, directa o indirectamente, se maximiza la verosimilitud de los datos dado el modelo.

Otros clasificadores Bayesianos relajan la restricción de independencia condicional impuesta en el naive Bayes para así modelar relaciones más complejas entre variables. Por ejemplo, en el naive Bayes aumentado a árbol (TAN) [7] se permite que las variables predictoras formen entre ellas una estructura de árbol. El modelo de clasificador Bayesiano k-dependiente [15] permite modelar un conjunto de dependencias entre variables predictoras más amplio. En éste modelo se amplía la estructura del clasificador naive Bayes permitiendo que cada variable predictora tenga hasta un máximo de k variables predictoras como padres. Otro ejemplo de clasificador Bayesiano más complejo son los clasificadores basados en redes Bayesianas múltiplemente conectadas [13, 19]. En estos últimos modelos la clasificación se efectúa por medio de modelos gráficos probabilísticos en los que no se restringen las dependencias entre variables. Todos estos clasificadores Bayesianos son considerados, generalmente, modelos generativos.

3. Aprendizaje discriminativo de clasificadores Bayesianos

En esta sección nos centraremos en el aprendizaje discriminativo de clasificadores Bayesianos. Para ello, primero debemos introducir el aprendizaje discriminativo de los parámetros. En nuestro caso, éste se basa en el algoritmo TM [5] adaptado para el aprendizaje discriminativo de clasificadores Bayesianos por Santafé y col. [17]. En el presente artículo únicamente se pretende dar una idea intuitiva del funcionamiento del algoritmo TM, por lo que únicamente se plantea la estructura general del algoritmo. Para más detalle consultar [16, 17]. Posteriormente nos centraremos más en el uso del algoritmo TM para el aprendizaje de la estructura del clasificador.

3.1. El algoritmo TM de Edwards y Lauritzen

El algoritmo TM, es un algoritmo general que permite maximizar la verosimilitud condicional para aquellos modelos, como es el caso de las redes Bayesianas, en los que es más sencillo maximizar la verosimilitud conjunta. Se basa en una aproximación a la función de verosimilitud condicional a través de la función de verosimilitud conjunta. Para introducir el desarrollo del algoritmo TM, denotemos como:

$$\begin{aligned} l(\boldsymbol{\theta}) &= \log f(\boldsymbol{x}, c|\boldsymbol{\theta}) \\ l_{\boldsymbol{x}}(\boldsymbol{\theta}) &= \log f(\boldsymbol{x}|\boldsymbol{\theta}) \\ l^{\boldsymbol{x}}(\boldsymbol{\theta}) &= \log f(c|\boldsymbol{x}, \boldsymbol{\theta}) \end{aligned}$$

las funciones de verosimilitud conjunta, marginal y condicional respectivamente, donde $\boldsymbol{\theta}$ es el conjunto de parámetros para la función de distribución conjunta. Si tenemos en cuenta que la verosimilitud condicional puede ser expresada en función de la marginal y la conjunta:

$$l^{\boldsymbol{x}}(\boldsymbol{\theta}) = l(\boldsymbol{\theta}) - l_{\boldsymbol{x}}(\boldsymbol{\theta}) \tag{1}$$

podemos aproximar la función de verosimilitud condicional desarrollando la Ecuación 1 en serie de Taylor de primer orden en el punto $\boldsymbol{\theta}_r$. De esta forma, omitiendo los términos constantes respecto a $\boldsymbol{\theta}$, obtenemos:

$$l^{\boldsymbol{x}}(\boldsymbol{\theta}) \approx q(\boldsymbol{\theta}|\boldsymbol{\theta}_r) = l(\boldsymbol{\theta}) - \boldsymbol{\theta}^T \dot{l}_{\boldsymbol{x}}(\boldsymbol{\theta}_r) \tag{2}$$

donde $\dot{l}_{\boldsymbol{x}}(\boldsymbol{\theta}_r)$ es la derivada de $l_{\boldsymbol{x}}(\boldsymbol{\theta})$ en el punto $\boldsymbol{\theta}_r$. Por otra parte tenemos que, bajo condiciones regulares de la función de verosimilitud, podemos afirmar que $\dot{l}_{\boldsymbol{x}}(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}\{\dot{l}(\boldsymbol{\theta})|\boldsymbol{x}\}$ (ver [5, 17]). De esta forma la aproximación a la verosimilitud condicional mediante la función $q(\boldsymbol{\theta}|\boldsymbol{\theta}_r)$ se puede expresar en términos que sólo dependen de la verosimilitud conjunta:

$$q(\boldsymbol{\theta}|\boldsymbol{\theta}_r) = l(\boldsymbol{\theta}) - \boldsymbol{\theta}^T E_{\boldsymbol{\theta}}\{\dot{l}(\boldsymbol{\theta})|\boldsymbol{x}\} \tag{3}$$

Como las derivadas de la verosimilitud condicional y de la función $q(\boldsymbol{\theta}|\boldsymbol{\theta}_r)$ con respecto a $\boldsymbol{\theta}$ tienen el mismo valor en el punto $\boldsymbol{\theta}_r$, podemos maximizar $l^{\boldsymbol{x}}(\boldsymbol{\theta})$ maximizando $q(\boldsymbol{\theta}|\boldsymbol{\theta}_r)$.

Por consiguiente, la obtención de los parámetros que maximicen la verosimilitud condicional se convierte en un proceso iterativo que consta de dos pasos:

- Paso T: dado el actual valor de $\boldsymbol{\theta}_r$ construir la función $q(\boldsymbol{\theta}|\boldsymbol{\theta}_r)$
- Paso M: obtener el siguiente conjunto de parametros que maximicen la verosimilitud condicional

$$\boldsymbol{\theta}_{r+1} = \arg \max_{\boldsymbol{\theta}} \quad q(\boldsymbol{\theta}|\boldsymbol{\theta}_r)$$

Particularizando el algoritmo TM para las distribuciones de probabilidad pertenecientes a la familia exponencial (familia a la que pertenecen las redes Bayesianas), deberiamos expresar las funciones de verosimilitud conjunta y condicional de la siguiente forma:

$$l(\boldsymbol{\theta}) = \boldsymbol{\alpha}^T u(c, \boldsymbol{x}) + \boldsymbol{\beta}^T v(\boldsymbol{x}) - \psi(\boldsymbol{\alpha}, \boldsymbol{\beta}) \quad (4)$$

$$l^x(\boldsymbol{\theta}) = \boldsymbol{\alpha}^T u(c, \boldsymbol{x}) - \psi^x(\boldsymbol{\alpha}) \quad (5)$$

siendo:

$$\psi(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \log \int exp\{\boldsymbol{\alpha}^T u(c, \boldsymbol{x}) + \boldsymbol{\beta}^T v(\boldsymbol{x})\} \mu(dc|\boldsymbol{x}) \mu(d\boldsymbol{x})$$

$$\psi^x(\boldsymbol{\alpha}) = \log \int exp\{\boldsymbol{\alpha}^T u(c, \boldsymbol{x})\} \mu(dc|\boldsymbol{x})$$

y donde $\boldsymbol{\alpha}$ denotan los parámetros del modelo condicional que junto a $\boldsymbol{\beta}$ forman los parámetros del modelo conjunto – $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta})$ –. Por otra parte $\boldsymbol{\mathcal{U}} = E_{\boldsymbol{\theta}}\{u(c, \boldsymbol{x})\}$ son los estadísticos suficientes del modelo condicional, que junto con $\boldsymbol{\mathcal{V}} = E_{\boldsymbol{\theta}}\{v(\boldsymbol{x})\}$ forman los estadísticos suficientes del modelo conjunto.

Por tanto, si desarrollamos las expresiones $E_{\boldsymbol{\theta}_r}\{\dot{l}(\boldsymbol{\theta}_r)|\boldsymbol{x}\}$ y $\dot{l}(\boldsymbol{\theta})$ de forma adecuada, obtenemos que el paso M de la $r+1$-ésima iteración del algoritmo requiere la actualización de los estadísticos suficientes del modelo condicionado y los parámetros del modelo conjunto de la siguiente forma:

$$\boldsymbol{u}_{r+1} = \boldsymbol{u}_r + \boldsymbol{u}_0 - E_{\boldsymbol{\theta}_r}\{\boldsymbol{\mathcal{U}}|\boldsymbol{x}\}$$

$$\boldsymbol{\theta}_{r+1} = \hat{\boldsymbol{\theta}}(\boldsymbol{u}_{r+1}, \boldsymbol{v}) \quad (6)$$

donde los estadísticos suficientes iniciales, $\boldsymbol{u}_0$ y $\boldsymbol{v}$, vienen dados por los estadísticos máximo verosímiles obtenidos directamente del conjunto de datos. Además, $\hat{\boldsymbol{\theta}}(\boldsymbol{u}_{r+1}, \boldsymbol{v})$ denota la estimación máximo verosímil de $\boldsymbol{\theta}$ obtenida de los estadísticos suficientes $\boldsymbol{u}_{r+1}$ y $\boldsymbol{v}$.

Puede darse el caso en el que la actualización de los parámetros $\boldsymbol{\theta}$ en una de las iteraciones del algoritmo de como resultado un conjunto de parámetros ilegal, o que el valor de la verosimilitud condicional decrezca en dos iteraciones consecutivas. Estas situaciones deben ser corregidas mediante una búsqueda local en la cual la actualización de $\boldsymbol{u}_{r+1}$ descrita en la Ecuación 6 se sustituye por:

$$\boldsymbol{u}_{r+1} = \boldsymbol{u}_r + \lambda(\boldsymbol{u}_0 - E_{\boldsymbol{\theta}_r}\{\boldsymbol{\mathcal{U}}|\boldsymbol{x}\}) \quad (7)$$

siendo $\lambda \in (0, 1)$ el valor que maximice la verosimilitud condicional.

3.2. El algoritmo TM estructural

En esta sección presentamos el aprendizaje estructural de redes Bayesianas usando el algoritmo TM. Éste es un algoritmo voraz de búsqueda en el cual se parte de una red Bayesiana vacía. En cada paso del algoritmo hay un conjunto de arcos candidatos que pueden ser añadidos a la red y que está formado por todos aquellos arcos cuya inclusión en la red no crea un ciclo en la estructura gráfica. Así pues, en cada iteración, todos los modelos obtenidos mediante la inclusión de cada uno de los arcos candidatos en el clasificador son evaluados en base a una medida de calidad o *score*. Aquel modelo con mayor *score* es seleccionado, se vuelve obtener los arcos candidatos para el nuevo modelo y se repite el proceso hasta que el valor del *score* no mejore. En concreto, el *score* utilizado en el algoritmo TM estructural es una adaptación de la métrica BIC [18]. La métrica BIC original utiliza el logaritmo de la verosimilitud conjunta penalizada con la complejidad del modelo. Sin embargo, y dado que nuestra intención es hacer un aprendizaje discriminativo, nuestra métrica BIC discriminativa (BICd) utilizará el logaritmo de la verosimilitud condicional en lugar de la conjunta:

$$BICd(D|\mathcal{M}) = \sum_{i=1}^{N} \log p(c^i|\boldsymbol{x}^i, \mathcal{M}) - \frac{\log N}{2} d \quad (8)$$

donde D es el conjunto de datos, $D = \{(\boldsymbol{x}^1, c^1), \ldots, (\boldsymbol{x}^N, c^N)\}$, N es el número de casos en el conjunto de datos, $\mathcal{M}$ es el modelo de clasificación Bayesiano que estamos evaluando y d es el número de parámetros de dicho modelo. Téngase en cuenta que, en nuestro caso, el modelo $\mathcal{M}$ es un modelo condicional (discriminativo) y por tanto el número de parámetros, d, también es el número de parámetros del modelo condicionado. Es decir, teniendo en cuenta la descripción del aprendizaje de parámetros mediante el algoritmo TM para modelos de la familia exponencial, el número de parámetros, d, se correspondería con la cardinalidad de $\boldsymbol{\alpha}$ en la Ecuación 5.

En el proceso de búsqueda estructural, cada vez que evaluamos un nuevo modelo candidato, los parámetros del modelo se aprenden mediante el algoritmo TM. De esta forma, el conjunto de parámetros del modelo maximizan la verosimilitud condicional. Sin embargo, dentro del proceso de aprendizaje estructural el modelo en sí es evaluado en base a la métrica BICd descrita anteriormente.

Para la aplicación de la búsqueda estructural basada en el algoritmo TM se requiere que todas las variables sean discretas y que no haya datos perdidos. Por tanto, puede que se requiera un pre-procesado de los datos antes de aprender el clasificador Bayesiano. No obstante, al estar el algoritmo TM basado en estadísticos suficientes obtenidos a partir del conjunto de datos, la restricción de datos completos puede relajarse en un futuro mediante el uso del algoritmo EM [3].

4. Experimentos

En esta sección presentamos una primera evaluación experimental que pretende ilustrar el comportamiento del aprendizaje discriminativo de clasificadores Bayesianos mediante el algoritmo TM estructural. Debido a la complejidad del proceso de la búsqueda estructural, y debido a que sólo contamos con una implementación del algoritmo TM capaz de aprender modelos hasta una complejidad de TAN, la búsqueda estructural también se limita a modelos TAN. Es decir, cada una de las variables predicitoras podrá tener como máximo dos padres: la variable clase y otra variable predictora. La evaluación experimental se realiza sobre algunas de las bases de datos utilizadas en [7], las cuales se han convertido en un estándar para la evaluación de clasificadores Bayesianos. Las bases de datos utilizadas son `Breast`, `Cleve`, `Crx`, `German`, `Heart`, `Hepatitis`, `Iris`, `Lymphography` y `Vote` procedentes del repositorio de UCI [1], y la base de datos `Corral` diseñada, entre otras, por Kohavi y John [11] explicitamente para evaluar métodos de selección de variables.

Tal y como se apuntaba anteriormente, el algoritmo TM, por el momento, no es capaz de manejar variables continuas ni casos perdidos. Dado que en las bases de datos utilizadas nos encontramos con ambas situaciones, ha sido necesario un pre-procesado de los datos. Por una parte todas las variables con valores continuos han sido discretizadas utilizando el método descrito en [4], el cual es una variante de la discretización de Fayyad e Irani [6]. Por otra parte, todos los casos de las bases de datos en los que aparezca algún caso perdido han sido eliminados y no han sido tenidos en cuenta para el aprendizaje y evaluación de los modelos.

En la búsqueda estructural, los parámetros de cada uno de los modelos candidatos son aprendidos mediante el algoritmo TM. El algoritmo TM es un proceso iterativo que requiere fijar una serie de parámetros. Para los experimentos realizados, el algoritmo TM tiene fijado como criterio de parada que el valor obtenido para la verosimilitud condicional en dos iteraciones consecutivas sea menor que 0,001. Por otra parte, a veces el conjunto de parámetros obtenido en una de las iteraciones del algoritmo TM no es válido y es necesario aplicar una búsqueda local para corregirlo (ver Ecuación 7). En nuestro caso, la búsqueda local se realiza buscando el mejor λ en el intervalo $(0, 1)$ con un incremento de 0,01.

Los modelos obtenidos mediante el algorit-

	TM Estructural	*NB–TM*	*TAN–TM*
Breast	95,02 ± 8,55	98,98 ± 0,74	95,46 ± 1,41
Cleve	82,42 ± 4,89	87,53 ± 4,72	87,85 ± 3,24
Iris	95,33 ± 3,40	95,33 ± 3,40	96,00 ± 2,49
German	73,30 ± 3,75	78,90 ± 4,00	84,00 ± 0,89
Corral	98,46 ± 3,08	90,61 ± 6,27	99,20 ± 1,60
Crx	86,06 ± 3,32	88,52 ± 1,59	89,59 ± 1,56
Lymphography	71,61 ±10,92	91,22 ± 3,49	98,98 ± 1,65
Hepatitis	87,50 ± 7,90	93,75 ± 5,56	100,00 ± 0,00
Vote	95,63 ± 2,45	98,39 ± 1,17	99,08 ± 0,86
Heart	85,56 ± 2,46	86,67 ± 4,44	81,75 ± 3,87

Cuadro 1: Precisión estimada obtenida en los experimentos para cada uno de los modelos y bases de datos

mo TM estructural son evaluados en base al porcentaje de bien clasificados y al logaritmo de su verosimilitud condicional. Este porcentaje de bien clasificados es obtenido mediante una validación cruzada en cinco hojas. El mismo procedimiento de validación ha sido utilizado anteriormente en la bibliografía para la evaluación tanto de clasificadores Bayesianos generativos [7] como discriminativos [8, 14, 17]. Además, los resultados obtenidos, tanto en el logaritmo de la verosimilitud condicional como en el porcentaje de bien clasificados, son comparados con otros modelos como son el naive Bayes y el TAN, en los que sus parámetros han sido aprendidos desde un punto de vista discriminativo mediante el algoritmo TM. La estructura del clasificador naive Bayes es fija, todas las variables predictoras son condicionalmente independientes dada la clase. Sin embargo, la estructura del modelo TAN depende de los datos y puede ser aprendida de diversas formas. En nuestro caso, la estructura del modelo TAN se aprende mediante el algoritmo de Friedman [7]. Es decir, aunque la estructura del modelo TAN es aprendida desde un punto de vista generativo, sus parámetros son aprendidos mediante un método discriminativo, el algoritmo TM.

En el Cuadro 1 se muestra el valor estimado del porcentaje de bien clasificados para los clasificadores considerados en los experimentos. El porcentaje estimado de bien clasificados para los modelos aprendidos con el algoritmo TM estructural es, por lo general, ligeramente inferior que el del mejor modelo aprendido (ya sea éste NB-TM o TAN-TM). Esto no quiere decir que definitivamente el modelo aprendido mediante el algoritmo estructural TM sea peor que el resto, ya que debemos tener en cuenta que en el modelo TM estructural no aparecen todas las variables predictoras y por tanto la complejidad del modelo puede verse reducida considerablemente, con las ventajas que ésto acarrea. Por consiguiente, en conjuntos de datos con gran número de variables y para los cuales el aprendizaje de un modelo con todas ellas puede ser realmente costoso, el algoritmo TM estructural puede contribuir a una simplificación del modelo sin tener porque sacrificar excesivamente la precisión en la clasificación. Además, esta simplificación del modelo también contribuye a una mayor facilidad para su interpretación. En la Figura 1 aparecen las estructuras de los clasificadores que han sido aprendidos mediante el algoritmo TM estructural utilizando las distintas bases de datos que se han tenido en cuenta para los experimentos. Nótese que, la reducción del número de variables con respecto los modelos naive Bayes y TAN aprendidos de las bases de datos con todas las variables, en la mayoría de los casos, es más que considerable (ver Cuadro 2).

El Cuadro 3 presenta los valores del logaritmo de la verosimilitud condicional para los modelos aprendidos de cada una de las bases

de datos. En estos resultados, se vuelve a constatar que los valores obtenidos para los modelos aprendidos con el algoritmo TM estructural son, por lo general, ligeramente inferiores que los obtenidos por el resto de modelos. Esto es un hecho razonable ya que para la búsqueda estructural no se utiliza directamente como *score* la verosimilitud condicional sino una versión penalizada de ésta (ver ecuación 8). Por tanto en la búsqueda estructural no se busca la estructura que maximice la verosimilitud condicional sino un equilibrio entre máxima verosimilitud condicional y complejidad de la estructura.

	Variables Base Datos	*Variables Modelo Discriminativo*
Breast	9 + Clase	2 + Clase
Cleve	13 + Clase	5 + Clase
Iris	4 + Clase	1 + Clase
German	20 + Clase	4 + Clase
Corral	6 + Clase	5 + Clase
Crx	15 + Clase	5 + Clase
Lymphography	18 + Clase	4 + Clase
Hepatitis	19 + Clase	2 + Clase
Vote	16 + Clase	3 + Clase
Heart	13 + Clase	4 + Clase

Cuadro 2: Número de Variables en cada una de las bases de datos utilizadas y número de variables seleccionadas por el algoritmo TM estructural.

5. Conclusiones y trabajo futuro

En el presente trabajo hemos presentado el algoritmo TM estructural, un nuevo método discriminativo para el aprendizaje de la estructura y los parámetros de clasificadores Bayesianos. Este nuevo algoritmo está basado en el algoritmo TM [5] que ha sido recientemente adaptado para el aprendizaje discriminativo de parámetros en clasificadores Bayesianos por Santafé y col. [17]. Mediante experimentos con bases de datos estándares se ha visto que, a pesar de que el rendimiento de los clasificadores obtenidos mediante el algoritmo TM estructural no es mejor que el de otros clasificadores con los que se los ha comparado, la reducción de la complejidad del modelo hace que éstos sean computacionalmentes más asequibles de manejar. Por otra parte, la reducción en la complejidad también puede facilitar interpretabilidad de los modelos.

	TM Estructural	*NB−TM*	*TAN−TM*
Breast	−68,94	−22,71	−10,41
Cleve	−100,45	−93,24	−82,27
Iris	−23,13	−13,85	−16,57
German	−445,71	−456,81	−378,12
Corral	−3,60	−25,17	−3,28
Crx	−196,85	−177,90	−173,71
Lymphography	−58,85	−28,06	−13,17
Hepatitis	−16,80	−9,27	−2,61
Vote	−41,08	−13,66	−13,88
Heart	−98,57	−86,11	−73,66

Cuadro 3: Valor del logaritmo de la verosimilitud condicional obtenido en los experimentos para cada uno de los modelos y bases de datos.

Como trabajo futuro se puede destacar una evaluación más exhaustiva del algoritmo TM estructural sobre un mayor número de bases de datos. Además, los autores ven factible la implementación de una versión paralela del algoritmo que permita hacer la búsqueda estructural de forma más rápida, así como la extensión de la busqueda estructural a estructuras más complejas que TAN. Por otra parte, uno de los inconvenientes del algoritmo es que no puede tratar con datos perdidos. Este problema puede resolverse utilizando un método que combine los algoritmos EM y TM.

6. Agradecimientos

Este trabajo ha sido financiado por el Ministerio de Ciencia y Tecnología mediante la beca TIC2001-2973-C05-03, por el Gobierno Vasco con los proyectos ETORTEK-BIOLAN y SAIOTEK S-PE04UN25, por la Universidad del País Vasco mediante la beca 9/UPV 00140.226-15334/2003 y por el Gobierno de Navarra con una beca predoctoral concedida al primer autor de este artículo.

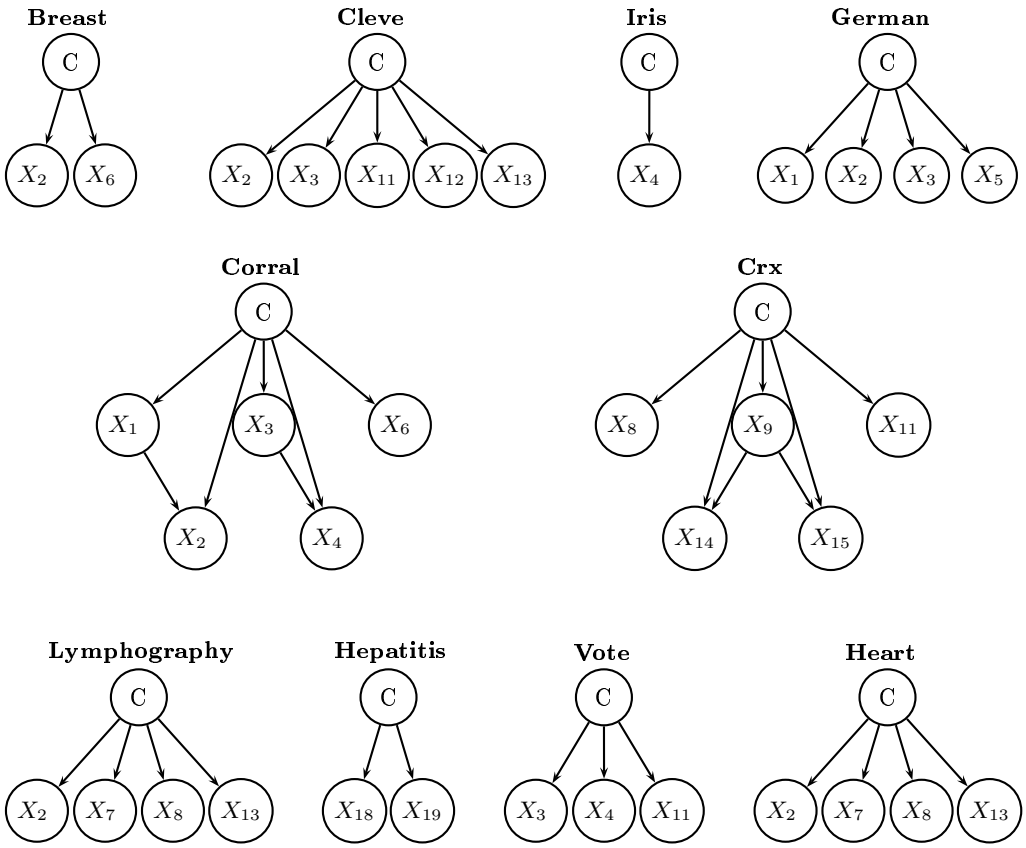

Figura 1: Estructuras de los clasificadores obtenidos mediante el algoritmo TM estructural en cada una de las bases de datos utilizadas.

Referencias

[1] C.L. Blake y C.J. Merz. UCI repository of machine learning databases. http://www.ics.uci.edu/~mlearn, 1998.

[2] G. F. Cooper y E. Herskovits. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9:309–347, 1992.

[3] A. P. Dempster, N. M. Laird, y D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society B*, 39:1–38, 1977.

[4] J. R. Dougherty, R. Kohavi, y M. Sahami. Supervised and unsupervised discretization of continuous features. En *Proceedings of the 12th International Conference on Machine Learning*, pages 194–202, 1995.

[5] D. Edwards y S. L. Lauritzen. The TM algorithm for maximising a conditional likelihood function. *Biometrika*, 88(4):961–972, 2001.

[6] U. Fayyad y K. Irani. Multi-interval discretization of continuous-valued attributes for classification learning. En *Proceedings of the 13th International Joint Conference on Artificial Intelligence*, pages 1022–1027, 1993.

[7] N. Friedman, D. Geiger, y M. Goldszmidt. Bayesian networks classifiers. *Machine Learning*, 29:131–163, 1979.

[8] R. Greiner, W. Zhou, X. Su, y B. Shen. Structural extension to logistic regression: Discriminative parameter learning of belief net classifiers. *Machine Learning*, 2005. Aceptado.

[9] D. Grossman y P. Domingos. Learning Bayesian network classifiers by maximizing conditional likelihood. En *Proceedings of the 21st International Conference on Machine Learning*, pages 361–368, 2004.

[10] D. Heckerman. A tutorial on learning with Bayesian networks. Informe Interno MSR-TR-95-06, Microsoft Research, 1995.

[11] R. Kohavi y G. H. John. Wrappers for feature subset selection. *Artificial Intelligence*, 97(1-2):273–324, 1997.

[12] M. Minsky. Steps toward artificial intelligence. *Transactions on Institute of Radio Engineers*, 49:8–30, 1961.

[13] J. Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, 1988.

[14] T. Roos, H. Wettig, P. Grünwald, P. Myllymäki, y H. Tirri. On discriminative Bayesian network classifiers y logistic regression. *Machine Learning*, 2005. Aceptado.

[15] M. Sahami. Learning limited dependence bayesian classifiers. En *Proceedings of the 2n International Conference on Knowledge Discovery and Data Mining*, pages 335–338, 1996.

[16] G. Santafé, J. A. Lozano, y P. Larrañaga. El algoritmo TM para clasificadores Bayesianos. Informe Interno EHU-KZAA-IK-2/04, Universidad del País Vasco, 2004.

[17] G. Santafé, J. A. Lozano, y P. Larrañaga. Discriminative learning of Bayesian network classifiers via the TM algorithm. En *Proceedings of the 8th European Conference on Symbolic and Quantitative Approaches to Reasonig with Uncertainty*, 2005. Aceptado.

[18] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6:461–464, 1978.

[19] B. Sierra y P. Larrañaga. Searching for the optimal Bayesian network in classification tasks by genetic algorithms. En *Proceedings of the Workshop on Uncertainty Processing*, pages 144–155, 1997.