Clasificadores bayesianos en cadena utilizando el algoritmo Rebane-Pearl y el clasificador TAN para la clasificación multidimensional

Dulce Jazmín Navarrete Arias, Manuel Bahena Gómez y Osvaldo Navarro Guzmán
dj.navarrete, mbahena, onavarro {@ccc.inaoep.mx}

Instituto Nacional de Astrofísica, Óptica y Electrónica
Departamento de Ciencias Computacionales

Resumen La clasificacion multidimensional consiste en asignar a cada objeto de un conjunto un grupo de clases. Los enfoques actuales para resolver este problema presentan desventajas como: alta complejidad computacional e ignorar dependencia entre las clases y entre los atributos. Uno de estos métodos consiste en generar una cadena de clasificadores Bayesianos generada por medio de el algoritmo Chow-Liu. En este trabajo se proponen dos mejoras con la finalidad de evitar las desventajas antes mecionadas. Las propuestas consisten en implementar un clasificador TAN que incluye las relaciones entre los atributos, además de consideras las dependencias entre clases mediente el algoritmo de Rebane-Pearl.

1. Introducción

La clasificación tradicional consiste en asignar a cada objeto de un conjunto una sola clase. A diferencia de esto, el objetivo de la clasificación multidimensional (MDC) es asignar a cada objeto un conjunto de diferentes clases. De acuerdo con [1], esta tarea es importante ya que muchos problemas notables pueden ser vistos como un caso de clasificación multidimensional, e.g., clasificación de textos (asignar un documento a varios temas), selección de medicamentos contra el VIH (determinar el conjunto óptimo de drogas), etcétera.

Existen dos opciones principales para resolver el problema de la clasificación multidimensional: *binary relevance* y *label power-set* [1]. La opción de *binary relevance* consiste en transformar el problema de clasificación multidimensional en x problemas de clasificación binaria, uno para cada clase $C_1, C_2, ..., C_d$. Posteriormente, se entrena un clasificador de manera independiente para cada clase y los resultados se combinan para determinar el conjunto de clases.

Las ventajas del método de *binary relevance* son su baja complejidad computacional y que por medio de éste se puedan aplicar técnicas de clasificación

directamente. Sin embargo, este enfoque no considera las posibles relaciones entre las clases, lo cual afecta negativamente la precisión de la clasificación general.

La opción de *label power set* transforma el problema de clasificación multidimensional en uno de clasificación binaria, definiendo una variable que consiste en un conjunto de clases, la cual puede tomar todas las posibles combinaciones de los valores de las clases originales. En contraste con la opción anterior, este método considera las relaciones entre las clases. Sin embargo, conforme el número de clases del problema aumenta, el tamaño de la variable aumenta exponencialmente.

En [1], se propone un método para resolver el problema de MDC que combina las dos opciones anteriores, aprovechando las cualidades de cada una. Este enfoque consiste en dos etapas principales:

- Obtener una estructura de dependencias entre las clases.
- Construir una cadena de clasificadores por cada nodo de la estructura obtenida.

Éste método considera información mutua entre las clases para obtener la estructura inicial, sin embargo, no define dependencias entre las clases, lo cual podría aumentar el desempeño del clasificador. Además, en este enfoque no se consideran las posibles relaciones existentes entre los atributos de la estructura de dependencias, lo cual también podría aumentar la precisión de la clasificación.

En este trabajo se proponen 2 mejoras al método de *bayes chains* [1], con la finalidad de aumentar la precisión de la clasificación. Una de ellas, es implementar el algoritmo Rebane-Pearl para ayudar a definir las direcciones de la estructura de independencias obtenida en el método original. La segunda modificacion consiste en utilizar el clasificador TAN en cada nodo de la estructura para considerar las posibles relaciones entre los atributos.

El resto de este reporte está estructurado de la siguietne manera. En la sección 2 se expone el trabajo relacionado; en la sección 3, se describe la metodología; en la sección 4, se resumen los experimentos y resultados obtenidos; finalmente, en la sección 5, se muestran las conclusiones y el trabajo futuro.

2. Trabajo relacionado

En esta sección se describen brevemente, algunos trabajos relacionados que proponen diferentes métodos relacionados con la clasificación multidimensional.

En el artículo [2] se describen algunas ventajas del *binary relevance* sobre otros métodos más sofisticados,especialmente en términos de tiempo de procesamiento. Además, se propone un método de clasificación multidimensional basado en *binary relevance* que modela las correlaciones entre las clases y mant¡ene

una complejidad computacional aceptable. En [1], se presenta un método que combina las cualidades de las cadenas de clasificadores y las redes Bayesianas para la clasificación multidimensional. Este método consiste en dos fases. En la primera fase, una red Bayesiana que representa las relaciones de dependencia entre las variables de clase es generada a partir de los datos. En la segunda fase, se construye una serie de clasificadores en cadena, de tal forma que cada orden de las variables de clase en la cadena es consistente con la red Bayesiana.

3. Metodología y desarrollo

El método propuesto extiende el algoritmo de Cadenas de clasificadores Bayesianos (CCB), implementando dos mejoras. Estas modificaciones consisten en implementar el algoritmo Rebane-Pearl y el clasificador TAN. A continuación se descrbe el método del cual partimos, y las dos modificaciones propuestas.

Según [1], el método de Cadenas de clasificadores Bayesianos se puede resumir en 3 etapas. Dado un problema de clasificación multidimensional con d clases:

1 Construir un árbol no dirigido que se aproxime a la estructura de dependencias entre las clases (mediante el algoritmo Chow-Liu).
2 Crear d órdenes para los clasificadores en cadena, tomando cada clase como la raíz del árbol y asignando el resto de los nodos en el orden correspondiente.
3 Para cada clasificador en cada cadena se construye un clasificador Naïve Bayes con la clase C_i como raíz; el padre de C_i, i.e. $pa(C_i)$, junto con el vector de atributos x serán lás hojas.

Para clasificar una nueva instancia se combina la salida de las d cadenas (ensamble), y se utiliza un esquema de votación simple. En la figura 1 se muestra un ejemplo gráfico del modelo.

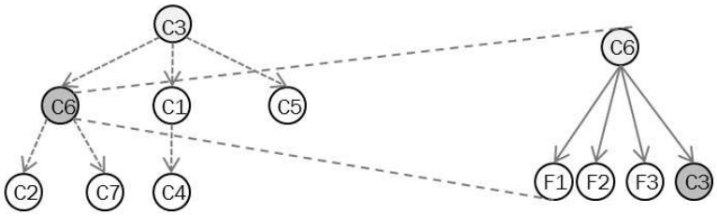

Figura 1. Ejemplo de una cadena de clasificadores Bayesianos.

Como se describe en [3], una forma de darle direcciones al "esqueleto" aprendido con el algoritmo de Chow-Liu, es mediante pruebas de independencias no

sólo entre dos variables, sino entre grupos de tres variables o tripletas. Mediante este esquema se genera un algoritmo que aprende poliárboles, ya que al asignar las direcciones puede ser que la estructura generada sea un árbol o un poliárbol (en realidad, un árbol es un caso especial de poliárbol).

Cuando se tienen atributos dependientes, una forma de considerar estas dependencias es extendiendo la estructura básica del clasificador Bayesiano simple agregando arcos entre dichos atributos. TAN agrega una estructura de árbol entre los atributos, de forma que se tienen en principio ´´pocas" conexiones y no aumenta demasiado la complejidad de la estructura. En algunos dominios, la precisión de la clasificación aumenta utilizando TAN. La figura 2 ilustra un ejemplo de un clasificador TAN

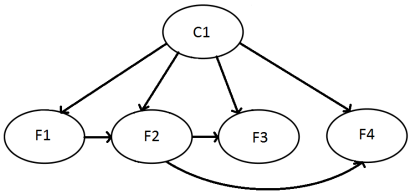

Figura 2. Ejemplo clasificador TAN.

El algoritmo parte de una estructura de dependencias sin direcciones. Después se determinan las direcciones de los arcos utilizando pruebas de dependencia entre tripletas de variables. Dadas tres variables, existen 3 casos posibles:

1 Arcos secuenciales: $X \rightarrow Y \rightarrow Z$.
2 Arcos divergentes: $X \leftarrow Y \rightarrow Z$.
3 Arcos convergentes: $X \rightarrow Y \leftarrow Z$.

Los primeros dos casos son indistinguibles con base en pruebas de independencia; es decir, son equivalentes. En ambos, X y Z son independientes dado Y. Pero el tercero es diferente, ya que las variables X y Z son marginalmente independientes. Este tercer caso lo podemos usar para determinar entonces las direcciones de los dos arcos que unen estas tres variables, y a partir de éstos, es posible encontrar las direcciones de otros arcos utilizando pruebas de independencia. De acuerdo a lo anterior, se establece el siguiente algoritmo para aprendizaje de poliárboles.

1 Obtener el esqueleto utilizando el algoritmo de Chow-Liu.
2 Recorrer la red hasta encontrar una tripleta de nodos que sean convergentes, donde la variable a la que apuntan los arcos la llamaremos *nodo multipadre*.
3 A partir de un nodo multipadre, determinar las direcciones de otros arcos utilizando la prueba de dependencia de tripletas, hasta donde sea posible (base causal).

4 Repetir 2-3 hasta que no se puedan descubrir más direcciones.
5 Si quedan arcos sin direccionar, utilizar semántica externa para obtener su dirección.

La figura 3 muestra un ejemplo de una estructura de poli-árbol semi-dirigida.

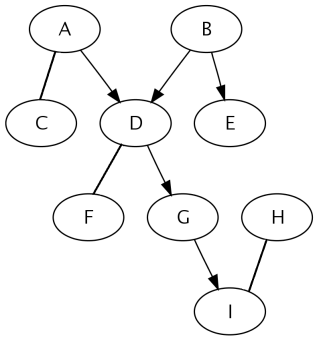

Figura 3. Ejemplo de una estructura semi-dirigida generada por el algoritmo de Pearl-Rebane.

El método que se propone puede dividirse en 7 etapas:

- Generación de estructura de dependencias.
- Direccionamiento de la estructura.
- Construcción de cadenas de clasificadores.
- Clasificación.

3.1. Generación de estructura de dependencias

A través del algoritmo de Chow-Liu se obtiene el esqueleto de una red Bayesiana con estructura de árbol sin direcciones, para las clases del conjunto de datos. En la figura 4 se muestra una estructura ejemplo generada por Chow-Liu.

3.2. Direccionamiento de la estructura

En esta etapa se parte del esqueleto (estructura sin direcciones) obtenido en la sección anterior y por medio del algoritmo Rebane-Pearl se determinan algunas direcciones de los arcos en la estructura. Si todos los arcos están direccionados, se usa la estructura tal y como está. De otra forma, para determinar las direcciones de los arcos que quedaron sin direccionar se siguen los siguientes criterios:

a Si existe *un solo arco* sin direccionar, se contruyen dos órdenes para la estructura: uno definiendo la dirección del arco hacia arriba y el otro direccionandolo hacia abajo.

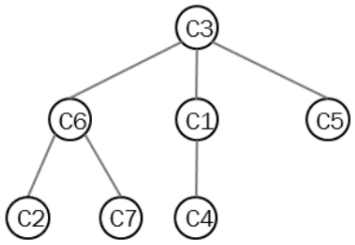

Figura 4. Ejemplo de una estructura generada por el algoritmo de Chow-Liu.

b Si existen *dos o más arcos* sin direccionar se seleccionan tres ordenes:
- Todas las direccionar hacia arriba
- Todas las direcciones hacia abajo
- Direcciones intercaladas (hacia arriba y hacia abajo)

3.3. Construcción de cadenas de clasificadores

Para la construcción cada una de las posibles cadenas, se siguen los siguientes pasos:

- Se recorre la estructura buscando los posibles nodos raíces, i.e. aquellos nodos hacia los cuales ningún nodo está apuntando.
- Se determinan los padres de cada nodo.
- Se genera la cadena de clasificadores, partiendo del nodo raíz hasta los nodos hoja, donde la clasificación se realizará en cada nodo, nivel por nivel.

En la figura 5 se muestra un ejemplo de cómo se genera una cadena de clasificadores Bayesianos.

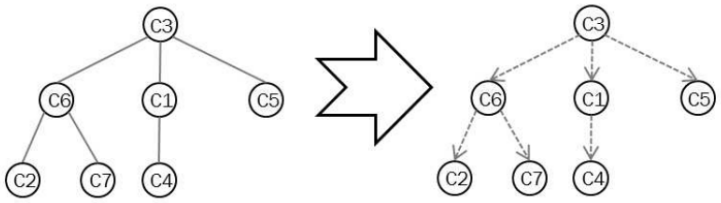

Figura 5. Ejemplo de generación de una cadena de clasificadores.

3.4. Clasificación

Una vez que se tiene la cadena de clasificadores, se procede a realizar la clasificación de los objetos. Para cada nodo en la cadena, se construye un clasificador TAN con la clase C_i como raíz; los padres de C_i, i.e. $pa(C_i)$ y todos los atributos forman las hijos en la estructura del clasificador.

La figura 6 muestra un ejemplo de la modificación descrita en esta sección.

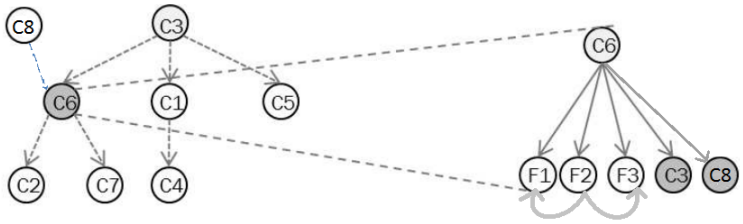

Figura 6. Ejemplo de una cadena de clasificadores Bayesianos con la modificación del TAN implementada.

Para clasificar una nueva instancia se combina la salida de las n cadenas, usando un esquema de votación simple.

4. Experimentos y resultados

En esta sección se describen los elementos utilizados durante la experimentación, además de los resultados obtenidos a través de ella. La Tabla 1 describe los recursos que se utilizaron para realizar los experimentos.

	Recurso	**Descripción**
Equipo	Intel(R) Core(TM)2 Duo, 2.00GHz, 4 GB de memoria RAM.	Utilizado en evaluación del baseline y el método propuesto.
Software	Eclipse-Java	Ambiente y lenguaje de desarrollo de la aplicación en la que se realizaron los experimentos y el baseline.
	Weka developer version 3.6.4	Librerias de los clasificadores utilizados en el baseline y los experimentos.
	MatLab R2010a	Aplicación en la que se generaron las estructuras de dependencias.

Tabla 1. Recursos utilizados en la experimentación

4.1. Conjuntos de datos

El método propuesto se evaluó con 4 conjuntos de datos multidimensionales. El rango de la dimensión de estos conjuntos va desde 6 hasta 20 clases. Las variables de clase de todos los conjuntos son binarias. Estos corpus se describen a detalle en la tabla 2

Conjunto	**No. de clases**	**No. de atributos**	**No. de instancias**
Emotions	6	72	593
Enron'	10	108	1186
Medical'	15	144	1500
Bibtex'	20	180	2130

Tabla 2. Descripción de los conjuntos de datos

El conjunto Emotions es el mismo que se utilizó en [1]. Los conjuntos Enron', Medical' y Bibtex' son subconjuntos de los conjuntos Enron, Medical y Bibtex del trabajo de investigación mencionado anteriormente. No se utilizaron estos conjuntos completos debido a limitaciones de recursos computacionales, particularmente la memoria RAM.

4.2. Baseline

El baseline con respecto al cual comparamos los resultados de los experimentos consiste en el método de Cadenas de clasificadores Bayesianos, i.e., aplicamos este clasificador a los conjuntos de datos antes mencionados, aplicamos luego el clasificador con las mejoras propuestas a los mismos conjuntos y comparamos sus resultados. El número de cadenas generadas para cada estructura fue el mismo que para el método propuesto.

4.3. Medidas de evaluación

El método elegido para la evaluación del rendimiento fue la validación cruzada con 10 pliegues. Además, para la comparación entre el método y el baseline se utilizaron las siguientes medidas de evaluación: *Mean accuracy* y *Global accuracy* (descritas en [1]). Estas medidas se muestran a detalle a continuación.

Mean accuracy Esta medida involucra las d variables de clase del conjunto de datos (precisión por etiqueta). La fórmula de esta medida se muestra a continuación:

$$\overline{Acc_d} = \frac{1}{d}\sum_{j=1}^{d} Acc_j = \frac{1}{d}\sum_{j=1}^{d}\frac{1}{N}\sum_{i=1}^{N}\delta(c'_{ij}, c_{ij}) \tag{1}$$

donde $\delta(c'_{ij}, c_{ij}) = 1$ si $c'_{ij} = c_{ij}$ y es 0 en otro caso. Note que c'_{ij}, denota al valor de salida de la clase C_j para el modelo del caso i, y que c_{ij} es su valor verdadero.

Global accuracy Esta medida involucra la variable de clase d-dimensional (precisión por instancia). La fórmula de la medida se muestra a continuación:

$$\overline{Acc} = \frac{1}{N}\sum_{i=1}^{N} \delta(c'_{ij}, c_{ij}) \tag{2}$$

donde $\delta(c'_{ij}, c_{ij}) = 1$ si y sólo si $c'_{ij} = c_{ij}$, y es 0 en otro caso. Entonces, se compara la coincidencia total de todos los valores de las clases predichas, con los valores reales de las clases.

4.4. Resultados

Los resultados obtenidos en los experimentos en cuanto a la medida mean accuracy se muestran la Tabla 3. De la misma forma, la Tabla 4 muestra la medida global accuracy obtenida en estas pruebas.

Conjunto	**CCB**	**CCB + Mejoras**
Emotions	0.77	0.80
Enron'	0.95	0.95
Medical'	0.99	0.99
Bibtex'	0.97	0.98

Tabla 3. Mean accuracy de los resultados

Conjunto	**CCB**	**CCB + Mejoras**
Emotions	0.26	0.30
Enron'	0.75	0.74
Medical'	0.93	0.93
Bibtex'	0.67	0.70

Tabla 4. Global accuracy de los resultados

4.5. Discusión de los resultados

Los resultados obtenidos en el primer experimento (mismo número de cadenas en baseline y método) superan o igualan la precisión obtenida en el baseline

en cuanto a la medida mean accuracy. Con respecto a la medida Global accuracy, las mejoras implementadas generan resultados superiores al baseline en 2 de los conjuntos de datos, obteniendo el mismo puntaje en un tercer conjunto y siendo ligeramente superado por el baseline en el último de ellos. Esto nos indica que el método propuesto no mejora significativamente el desempeño del enfoque original.

5. Conclusiones y trabajo futuro

En este trabajo de investigación se desarrollaron e implementaron dos mejoras al método de Cadenas de clasificadores Bayesianos, con la finalidad de aumentar la precisión de la clasificación. Los resultados obtenidos muestran una pequea mejora con respecto al método original. En el caso particular de Enron', los arcos generados quizás son inconsistentes con los datos, ya que al aplicar el algoritmo de Rebane-Pearl no se lograba generar un árbol con direcciones. De modo que para poder obtener algunas direcciones se amplió el umbral de test de independencia de Rebane-Pearl, el cual consiste en determinar la independiencia marginal entre los padres de un nodo, con lo cual podemos lograr tener direcciones pero se pierde consistencia en el modelo construido. Posiblemente, lo anterior se vio afectado por el tamaño limitado de los conjuntos de datos que se utilizaron.

Como trabajo futuro, se propone evaluar el método con conjuntos de datos de mayor tamaño y con distintas relaciones entre las clases. Además, evaluar este enfoque considerando todas las posibles cadenas consistentes con las direcciones generados por el algoritmo Rebane-Pearl podría darnos una idea más clara acerca del desempeño de este clasificador. Finalmente, sería interesante evaluar por separado las dos mejoras propuestas con el fin de evaluar el desempeño de cada una de ellas de forma independiente, tomando en cuenta que al tener un método más simple que otro y obteniendo similares resultados, es obvio que se preferirá tener un método mas sencillo.

6. Referencias

[1] *Sometido a revisión.* 2010. Bayes Chains Classifiers for Multidimensional Classification.

[2] Read, Pfahringer, Holmes & Frank. 2009. Classifier chains for multi-label classification. En *Proceedings ECML/PKDD*, páginas 254-269.

[3] Sucar. 2006. Aprendizaje automático: conceptos básicos y avanzados. Cap 6: Redes bayesianas. Páginas 77-98.