



ESCUELA TÉCNICA SUPERIOR DE INGENIEROS INFORMÁTICOS  
UNIVERSIDAD POLITÉCNICA DE MADRID

---

# Nuevos clasificadores Bayesianos multi-dimensionales. Aplicaciones a la eficiencia energética en la Industria 4.0

---

TRABAJO FIN DE MÁSTER  
MÁSTER UNIVERSITARIO EN INTELIGENCIA ARTIFICIAL

AUTOR: Santiago Gil Begué  
TUTORES: María Concepción Bielza Lozoya  
Pedro María Larrañaga Múgica

Julio, 2018



# **Nuevos clasificadores Bayesianos multi-dimensionales. Aplicaciones a la eficiencia energética en la Industria 4.0**

Santiago Gil Begué

## Tutores

---

Pedro Larrañaga Universidad Politécnica de Madrid

Concha Bielza Universidad Politécnica de Madrid

## Tribunal evaluador

---

Pedro Larrañaga (presidente) Universidad Politécnica de Madrid

Jesús Cardeñosa (secretario) Universidad Politécnica de Madrid

Alfonso Mateos (vocal) Universidad Politécnica de Madrid

## Financiación

---

Proyecto S2013/ICE-2845 Comunidad de Madrid

Proyecto PCD1610460364 Etxe-Tar S.A.



## **AGRADECIMIENTOS**

En primer lugar, quisiera dar mis agradecimientos a mis tutores Concha Bielza y Pedro Larrañaga por toda su excelente labor de dirección durante la realización de este trabajo. No sólo me han transmitido una fuente inmensa de sabiduría y ejemplo de investigación puntera, sino que también me han ofrecido una afabilidad única e interés en la promoción de mi formación mediante la participación en diferentes proyectos.

La colaboración en este trabajo con la empresa Etxe-Tar ha sido muy satisfactoria. En especial, agradezco a los directores de innovación Patxi Samaniego y Javier Díaz, con quienes he mantenido más contacto, por hacer este trabajo más atractivo y mantener en todo momento una disposición de colaborar y aprender juntos.

También me gustaría agradecer a los compañeros del grupo de investigación *Computational Intelligence Group* por su simpatía mostrada todos los días, así como a los compañeros del Máster por su amistad indiscutible. A Daki.

Agradezco también el apoyo de mi familia y de mis amigos. En especial a mis padres Manuel y María Dolores, a mis hermanos María del Mar y Manuel, y a Javier y María.

Este trabajo no hubiese sido posible sin el apoyo financiero de la Comunidad de Madrid mediante el proyecto S2013/ICE-2845-CASI-CAM-CM y el apoyo financiero de la empresa Etxe-Tar S.A. mediante el proyecto PCD1610460364.



# Nuevos clasificadores Bayesianos multi-dimensionales. Aplicaciones a la eficiencia energética en la Industria 4.0

## RESUMEN

El aprendizaje automático es una rama de la inteligencia artificial cuyo objetivo es inducir conocimiento automáticamente a partir de un conjunto de datos suministrados en forma de ejemplos. Dentro de este campo, la clasificación supervisada consiste en predecir el valor de habitualmente una variable clase en base al conocimiento de otras variables predictoras. Cuando son varias las variables que se quieren predecir de manera simultánea, el problema se denomina clasificación multi-dimensional. En este trabajo ponemos el foco de investigación sobre este tipo de problemas mediante una revisión exhaustiva de los clasificadores multi-dimensionales de redes Bayesianas, los cuales resuelven estos problemas de clasificación multi-dimensional. Además, investigamos nuevos modelos y como resultado proponemos un híbrido de estos clasificadores y los árboles de clasificación, así como un algoritmo *wrapper* para aprender este nuevo modelo desde un conjunto de datos. Un estudio experimental llevado a cabo sobre datos sintéticos muestra resultados alentadores en términos de precisión predictiva.

Por otra parte, en este trabajo abordamos mediante técnicas de aprendizaje automático el problema de la eficiencia energética en la Industria 4.0, la llamada cuarta revolución industrial que viene determinada por la introducción de las tecnologías digitales en la industria. La eficiencia energética se refiere a la práctica que tiene como objetivo la reducción del consumo de energía. Abordamos este problema mediante una colaboración con la empresa española del sector industrial Etxe-Tar S.A. en una de sus plantas de mecanizado. Para ello, intentamos predecir qué elementos activos de una máquina están consumiendo energía en un momento dado, lo que nos ayudará en un trabajo futuro a sincronizar de manera inteligente las puestas en marcha y paradas de estos elementos con su consecuente ahorro energético. La predicción la llevamos a cabo mediante los modelos de clasificación multi-dimensional investigados.

Finalmente, observamos que incorporar estos métodos desarrollados en una línea de producción en tiempo real requiere hacer frente a los potenciales cambios de concepto que surgen a lo largo del tiempo dentro de los flujos de datos. En este dominio en concreto, la huella energética de una máquina se deteriora con el tiempo, por lo que es necesario utilizar técnicas de aprendizaje adaptativo para hacer frente a estos cambios en los datos. Motivados por este hecho, realizamos un estudio exhaustivo de los métodos propuestos en la literatura para la clasificación de flujos de datos multi-dimensionales.



# **New multi-dimensional Bayesian classifiers. Applications to energy efficiency in Industry 4.0**

## **SUMMARY**

Machine learning is a branch of Artificial Intelligence whose objective is to automatically induce knowledge from a data set provided in the form of examples. Within this field, supervised classification consists of predicting the value of usually one class variable based on the knowledge of other feature variables. This problem is called multi-dimensional classification, when there are several variables to be predicted simultaneously. In this work, we put the focus of research on these type of problems through an exhaustive review of the multi-dimensional Bayesian network classifiers, which solve these multi-dimensional classification problems. In addition, we research new models and as a result we propose a hybrid of these classifiers and classification trees, as well as a wrapper algorithm to learn this new model from data. An experimental study carried out on randomly generated synthetic data shows encouraging results in terms of predictive accuracy.

On the other hand, in this work through machine learning techniques, we approach the problem of energy efficiency in Industry 4.0, the so-called fourth industrial revolution that is determined by the introduction of digital technologies in the industry. Energy efficiency refers to the practice that aims to reduce energy consumption. We approach this problem through a collaboration with the Spanish company of the industrial sector Etxe-Tar S.A. in one of its machining plants. For this, we try to predict which active elements of a machine are consuming energy at a given time, what will help us in a future work to intelligently synchronize the start-ups and stops of these elements with their consequent energy savings. The prediction is carried out through the researched multi-dimensional classification models.

Finally, we learn that incorporating these developed methods in a production line in real time requires facing the potential concept drifts that arise over time within the data streams. In this particular domain, the energy footprint of a machine deteriorates over time, so it is necessary to use adaptive learning techniques to deal with these changes in the data. Motivated by this fact, we conducted an exhaustive study of the state-of-the-art methods for the classification of multi-dimensional data streams.



# Índice

<b>1</b>	<b>Introducción</b>	<b>1</b>
1.1	Clasificación supervisada . . . . .	1
1.2	Problemas de clasificación multi-dimensional . . . . .	1
1.3	La cuarta revolución industrial . . . . .	3
1.4	El problema de la eficiencia energética . . . . .	5
1.5	Contribuciones . . . . .	8
1.6	Difusión científica . . . . .	9
<b>2</b>	<b>Clasificadores multi-dimensionales de redes Bayesiana</b>	<b>11</b>
2.1	Fundamentos . . . . .	11
2.1.1	Definición . . . . .	11
2.1.2	Familias de MBCs . . . . .	13
2.1.3	MBCs descomponibles . . . . .	14
2.2	Medidas de evaluación de rendimiento . . . . .	14
2.2.1	Medidas multi-etiqueta en la literatura . . . . .	15
2.2.2	Medidas multi-dimensionales en la literatura . . . . .	17
2.2.3	Medidas multi-dimensionales extendidas . . . . .	20
2.3	Complejidad en los MBCs . . . . .	22
2.3.1	Aprendizaje: cardinalidad del espacio de estructuras . . . . .	22
2.3.2	Inferencia: tratabilidad de las explicaciones más probables . . . . .	25
2.4	Aprendizaje a partir de un conjunto de datos . . . . .	29
2.4.1	Algoritmos <i>score-based</i> . . . . .	29
2.4.2	Algoritmos <i>constrained-based</i> . . . . .	31
2.4.3	MBCs tratables . . . . .	33
2.5	Aplicaciones en la literatura . . . . .	35
2.5.1	Problemas médicos . . . . .	35
2.5.2	Otras aplicaciones . . . . .	36
<b>3</b>	<b>Propuesta de un nuevo clasificador multi-dimensional en árbol</b>	<b>37</b>
3.1	Definición y contexto . . . . .	37

3.2	Algoritmo de aprendizaje <i>wrapper</i> . . . . .	38
3.3	Estudio experimental . . . . .	39
3.3.1	Resultados obtenidos . . . . .	40
<b>4</b>	<b>Aplicaciones en la Industria 4.0</b>	<b>43</b>
4.1	Definición del problema de clasificación . . . . .	43
4.2	Aproximación multi-etiqueta . . . . .	45
4.3	Aproximación multi-dimensional . . . . .	52
<b>5</b>	<b>Flujos de datos multi-dimensionales con cambio de concepto</b>	<b>57</b>
5.1	Definición y contexto . . . . .	57
5.2	Aproximaciones en la literatura . . . . .	58
5.2.1	Basadas en un único modelo . . . . .	59
5.2.2	Basadas en <i>ensemble</i> . . . . .	60
<b>6</b>	<b>Cierre del documento</b>	<b>63</b>
6.1	Conclusiones . . . . .	63
6.2	Discusión y trabajo futuro . . . . .	64
6.3	Evaluación personal . . . . .	66
<b>7</b>	<b>Bibliografía</b>	<b>67</b>
<b>Glosario</b>		<b>79</b>
<b>Anexos</b>		<b>81</b>
<b>A</b>	<b>Gestión del proyecto</b>	<b>83</b>
A.1	Planificación . . . . .	83
A.2	Metodología . . . . .	83
A.3	Herramientas utilizadas . . . . .	84

# Índice de figuras

1.1	Evolución de la industria y sus cuatro revoluciones industriales . . . . .	3
1.2	Los nueve pilares tecnológicos de la Industria 4.0 . . . . .	5
1.3	La potencia energética total de una vivienda mostrada a lo largo del tiempo refleja una serie de escalones en relación a los cambios de estado de diferentes electrodomésticos . . . . .	7
2.1	Ejemplo de la estructura de un MBC con sus tres subgrafos . . . . .	12
2.2	Ejemplos de estructuras gráficas en relación a diferentes familias de MBCs	13
2.3	Ejemplo de la estructura de un CB-MBC con sus dos subgrafos conexos máximas . . . . .	15
2.4	Ejemplo de moralización y poda de la estructura de un MBC. . . . .	27
3.1	Ejemplo de la estructura de un MBCTree . . . . .	37
3.2	Procedimiento del estudio experimental realizado con datos sintéticos para evaluar el modelo MBCTree . . . . .	40
4.1	Máquina de mecanizado sobre la que enfocamos el problema de eficiencia energética . . . . .	44
4.2	Evolución temporal de los estados de encendido y apagado de los tres elementos activos del problema de clasificación de eficiencia energética .	46
4.3	Detalle de las variables predictoras del problema de clasificación de eficiencia energética . . . . .	47
4.4	Estudio de las relaciones entre las variables predictoras del problema de eficiencia energética con la variable clase Spindle para la aproximación multi-etiqueta . . . . .	50
4.5	Distribución del conjunto de datos del problema de eficiencia energética en su versión multi-dimensional entre cada posible valor de las variables clase . . . . .	52
4.6	Estudio de las relaciones entre las variables predictoras del problema de eficiencia energética con la variable clase Spindle para la aproximación multi-dimensional . . . . .	54
5.1	Visualización de un cambio de concepto a lo largo del tiempo . . . . .	58
A.1	Planificación temporal del proyecto. . . . .	83



# Índice de tablas

1.1	Escenario típico de un problema de clasificación supervisada . . . . .	2
2.1	Equivalencia entre medidas de evaluación de rendimiento para problemas de clasificación multi-etiqueta y multi-dimensional . . . . .	21
2.2	Número de estructuras MBC para diferentes dimensiones . . . . .	25
2.3	Recopilación de los métodos propuestos en la literatura para aprender MBCs desde un conjunto de datos . . . . .	34
2.4	Problemas de clasificación multi-dimensional encontrados en la literatura que se han modelado con un MBC . . . . .	36
3.1	Comparación en términos de <i>global accuracy</i> de los modelos MBC y MBCTree siguiendo el experimento sobre datos sintéticos generados aleatoriamente . . . . .	41
4.1	Información mutua entre cada una de las variables del problema de eficiencia energética y las tres variables clase para la aproximación multi-etiqueta . . . . .	48
4.2	Distribución de los datos entre las diferentes configuraciones de clases para el problema de eficiencia energética en su aproximación multi-etiqueta	51
4.3	Rendimiento en términos de precisión predictiva de los modelos de clasificación entrenados para el problema de eficiencia energética en su aproximación multi-etiqueta . . . . .	52
4.4	Información mutua entre todas las variables del problema de eficiencia energética y las tres variables clase para la aproximación multi-dimensional	53
4.5	Rendimiento en términos de precisión predictiva de los modelos de clasificación entrenados para el problema de eficiencia energética en su aproximación multi-dimensional . . . . .	55
5.1	Recopilación de los métodos de clasificación de flujos de datos multi-dimensionales . . . . .	62



# Capítulo 1

## Introducción

### 1.1. Clasificación supervisada

El aprendizaje automático, también conocido como *machine learning*, es una rama de la inteligencia artificial cuyo objetivo es desarrollar técnicas que permitan a las máquinas aprender de manera automática. Más concretamente, se trata de crear programas capaces de generalizar comportamientos a partir de conjuntos de datos suministrados en forma de ejemplos. Es, por lo tanto, un proceso de inducción del conocimiento. El aprendizaje automático se ha aplicado en una amplia gama de dominios, tales como en medicina para el diagnóstico de enfermedades, en ciencia forense para identificar personas, en astrofísica para la clasificación de estrellas, en la predicción del tiempo, en bioinformática para la clasificación de secuencias de ADN, en la detección de fraude en el uso de tarjetas de crédito, en finanzas para el análisis del mercado de valores, en marketing para la segmentación de clientes similares, etc.

Dentro de los diferentes algoritmos de aprendizaje automático nos encontramos con los de clasificación supervisada, sobre los que ponemos el foco en este trabajo. En estos modelos predictivos, se desea conocer (predecir, clasificar) el valor de habitualmente una variable  $C$ , a la que se denomina clase, mediante el conocimiento de los valores de otras  $n$  variables,  $X_1, \dots, X_n$ , denominadas predictoras. Para ello, previamente se aprende un modelo desde un conjunto de datos etiquetado con  $N$  ejemplos. A este conjunto se la llama etiquetado porque se conocen las etiquetas  $c^{(i)}$  de la variable clase que queremos predecir en el futuro, gracias a las cuales la máquina es capaz de extraer conocimiento. En la Tabla 1.1 se muestra este escenario típico de clasificación supervisada, tal que después del último ejemplo etiquetado  $N$  deseamos predecir ejemplos no conocidos.

### 1.2. Problemas de clasificación multi-dimensional

En este trabajo estamos interesados en problemas de clasificación supervisada donde hay múltiples variables clase,  $C_1, \dots, C_d$ . El así llamado *problema de clasificación*

Tabla 1.1: Escenario típico de un problema de clasificación supervisada.

	$X_1$	$\dots$	$X_n$	C
$(\mathbf{x}^{(1)}, c^{(1)})$	$x_1^{(1)}$	$\dots$	$x_n^{(1)}$	$c^{(1)}$
$(\mathbf{x}^{(2)}, c^{(2)})$	$x_1^{(2)}$	$\dots$	$x_n^{(2)}$	$c^{(2)}$
$\dots$		$\dots$		$\dots$
$(\mathbf{x}^{(N)}, c^{(N)})$	$x_1^{(N)}$	$\dots$	$x_n^{(N)}$	$c^{(N)}$
$\mathbf{x}^{(N+1)}$	$x_1^{(N+1)}$	$\dots$	$x_n^{(N+1)}$	???

*multi-dimensional* consiste en encontrar una función  $h$  que asigna un vector de  $d$  variables clase  $\mathbf{c} = (c_1, \dots, c_d)$  a cada instancia dada por un vector de  $m$  variables predictoras  $\mathbf{x} = (x_1, \dots, x_m)$ :

$$h : \Omega_{X_1} \times \dots \times \Omega_{X_m} \rightarrow \Omega_{C_1} \times \dots \times \Omega_{C_d}$$

$$(x_1, \dots, x_m) \mapsto (c_1, \dots, c_d)$$

Asumimos que  $C_j$  es una variable discreta, para todo  $j \in \{1, \dots, d\}$ , tal que  $\Omega_{C_j}$  denota su espacio muestral e  $I = \Omega_{C_1} \times \dots \times \Omega_{C_d}$ , el espacio de configuraciones conjuntas de las variables clase. De manera análoga,  $\Omega_{X_i}$  es el espacio muestral de la variable predictora discreta  $X_i$ , para todo  $i \in \{1, \dots, m\}$ . Cuando todas las variables clase son binarias, i.e.,  $|\Omega_{C_j}| = 2$  para todo  $j \in \{1, \dots, d\}$ , el problema se denomina clasificación multi-etiqueta. Hay que notar que la clasificación multi-etiqueta es solamente una configuración particular de la clasificación multi-dimensional. Existen muchas contribuciones en la literatura a este paradigma multi-etiqueta. Dos revisiones actualizadas de las principales propuestas presentadas durante los últimos años son [Zhang and Zhou \[2014\]](#) y [Gibaja and Ventura \[2015\]](#).

Como establecieron [Bielza et al. \[2011\]](#), la clasificación multi-dimensional es un problema más difícil que aquella con una sola variable clase. El mayor problema es que hay un gran número de posibles combinaciones de los valores de las variables clase,  $|I|$ , y una dispersión habitual de los datos disponibles. En un escenario típico en el que una instancia  $\mathbf{x}$  se asigna a la combinación más probable (función de pérdida estándar 0-1), el objetivo es calcular  $\arg \max_{c_1, \dots, c_d} p(C_1 = c_1, \dots, C_d = c_d | \mathbf{x})$ . Se cumple que  $p(C_1 = c_1, \dots, C_d = c_d | \mathbf{x}) \propto p(C_1 = c_1, \dots, C_d = c_d, \mathbf{x})$ , lo que requiere asignar  $|I| \cdot |\Omega_{X_1} \times \dots \times \Omega_{X_m}|$  parámetros. En el caso con una única variable  $C$ ,  $|I|$  es solamente  $|\Omega_C|$  en vez de  $|\Omega_{C_1} \times \dots \times \Omega_{C_d}|$ . Además de ser un problema con una cardinalidad elevada, también es difícil estimar los parámetros requeridos a partir de un conjunto de datos (dispersos) en este espacio  $d$ -dimensional  $I$ . La factorización de esta distribución de probabilidad cuando se usa una red Bayesiana puede de alguna manera reducir

este número de parámetros, lo que ha sido estudiado con los llamados *clasificadores multi-dimensionales de redes Bayesiana*.

### 1.3. La cuarta revolución industrial

En los últimos años, se ha acuñado el término de Industria 4.0 o Industria conectada para referirse a un nuevo modo de coordinar los medios de producción. Este concepto hace referencia a la llamada cuarta revolución industrial que viene determinada por la introducción de las tecnologías digitales en la industria. En la Figura 1.1 se muestra la evolución de la industria desde su primera revolución, la cual ha estado marcada por una progresión en el aumento de la complejidad de las tecnologías que se han aplicado.

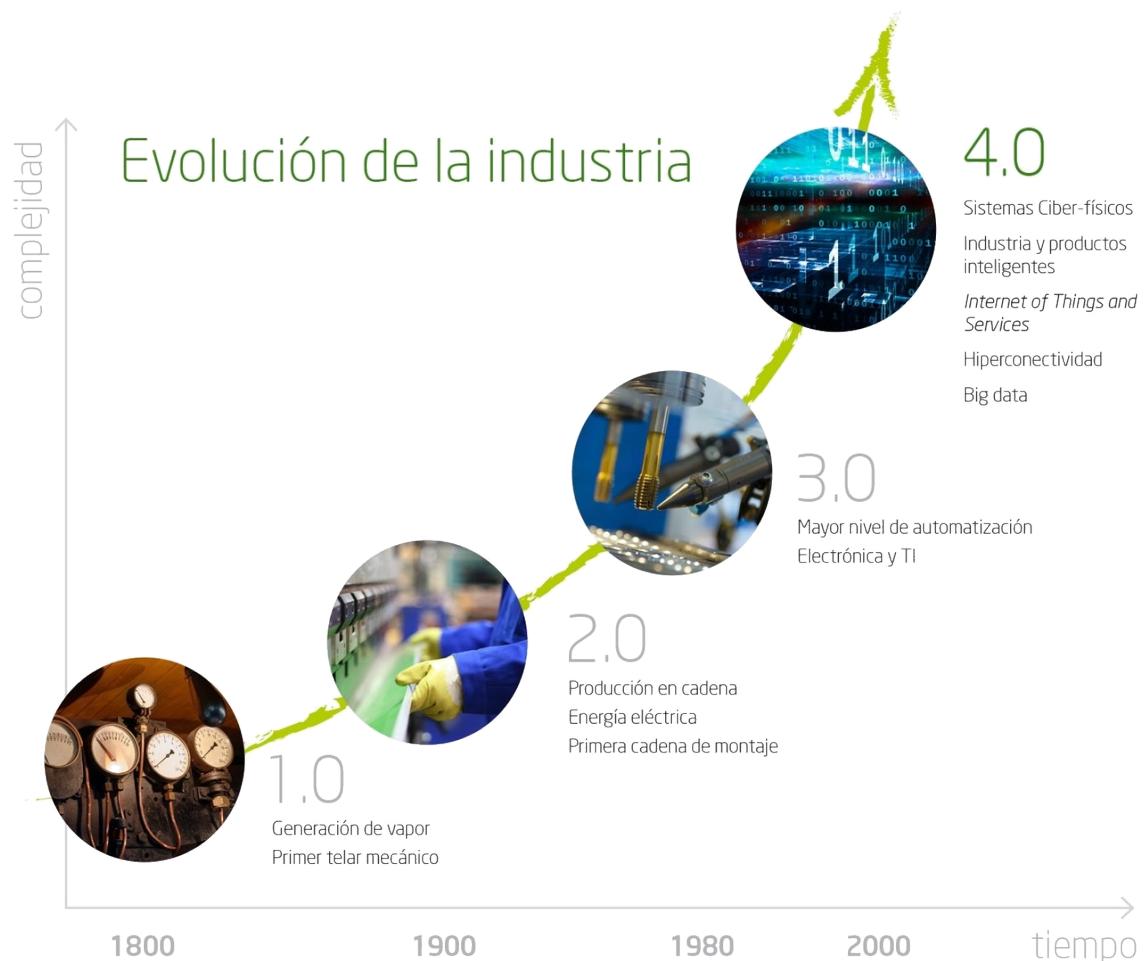


Figura 1.1: Evolución de la industria y sus cuatro revoluciones industriales. Fuente: Ministerio de Industria, Comercio y Turismo de España.

En esta transformación digital, diversos sensores, máquinas, piezas de trabajo y sistemas de la tecnología de la información se conectan a lo largo de la cadena de valor de un proceso industrial en las llamadas fábricas inteligentes. Estos sistemas conectados, también conocidos como sistemas ciber-físicos, pueden interactuar entre ellos para así

analizar datos y permitir la integración de la información digital con el mundo físico. Se requiere de esta manera que estos sistemas dispongan de una capacidad computacional suficiente para almacenar y procesar los datos [Hermann et al., 2016].

Gracias a esta recopilación y análisis de los datos entre máquinas, la Industria 4.0 permite conseguir procesos más rápidos, más flexibles y más eficientes para producir productos de mayor calidad a un costo reducido. Para ello, esta cuarta revolución toma su base en nueve avances tecnológicos (Figura 1.2) [Rüßmann et al., 2015]:

- **Big data.** Con la generación de grandes colecciones de datos, se dispondrá de la capacidad de tomar decisiones en tiempo real que mejoren la calidad de los productos y reducir los costes de producción.
- **Robots autónomos.** Los robots ya se están utilizando en muchas industrias, pero estos van a tener una mayor utilidad futura dadas las continuas mejoras en su autonomía, flexibilidad y cooperación. Estos robots tendrán la capacidad de interactuar entre ellos y con humanos, además de aprender de nosotros.
- **Simulación.** Las simulaciones aprovecharán los datos en tiempo real para reflejar el mundo físico en un modelo virtual, el cual puede incluir máquinas, productos y humanos. Permitirán la optimización de los procesos productivos en el mundo virtual antes de realizar cualquier cambio en el mundo físico.
- **Integración de sistemas.** Las empresas, sus departamentos, los clientes y los proveedores se volverán mucho más cohesivos, lo que permitirá conseguir cadenas de valor verdaderamente automatizadas.
- **Internet de las cosas.** Con el llamado internet de las cosas industrial, muchos más dispositivos se conectarán, creando una red que descentralice el análisis y la toma de decisiones además de permitir respuestas en tiempo real.
- **Ciberseguridad.** Las comunicaciones seguras y confiables, así como la administración sofisticada de identidades y accesos de máquinas y usuarios son esenciales ante la proliferación de las conexiones y el intercambio de datos.
- **La nube.** Muchas empresas ya utilizan aplicaciones basadas en la nube, pero con la Industria 4.0 muchas más tareas de producción requerirán un mayor intercambio de datos entre las empresas. Además, el rendimiento de las tecnologías en la nube mejorará, logrando tiempos de reacción de solo varios milisegundos, y consiguiendo de esta manera habilitar más servicios para sistemas de producción basados en datos y en la nube.
- **Producción aditiva.** La producción aditiva, como la impresión 3D, adquirirá un papel importante en el prototipado y fabricación de productos personalizados.

- **Realidad aumentada.** Esta tecnología se encuentra todavía en sus comienzos, pero en el futuro se le dará un mayor uso en la industria para proporcionar información en tiempo real a los trabajadores y así mejorar la toma de decisiones y los procedimientos de trabajo.

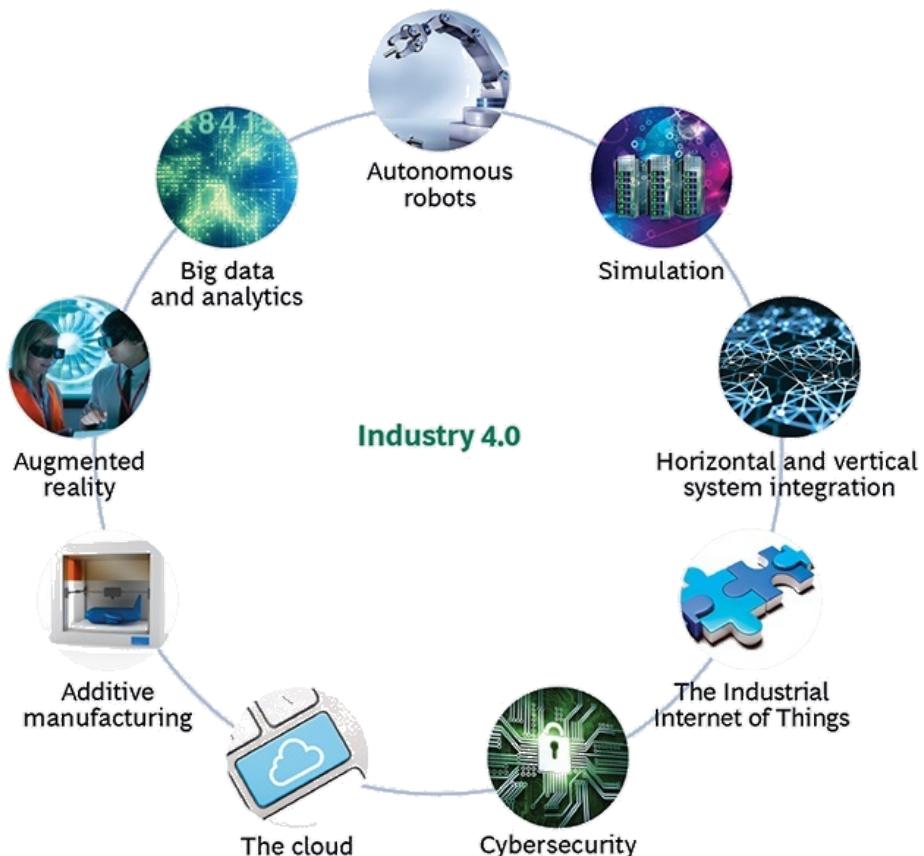


Figura 1.2: Los nueve pilares tecnológicos de la Industria 4.0. Fuente: [Rüßmann et al. \[2015\]](#).

En este trabajo nos centraremos en la utilización de la minería de datos y el aprendizaje automático para abordar el problema de la eficiencia energética en una industria de manufactura española. De los nueve avances tecnológicos comentados en la Industria 4.0, nosotros nos apoyaremos especialmente en el *big data* mediante la monitorización constante de las líneas de producción de la industria y en la nube para podernos aprovechar de estos datos directamente desde nuestro grupo de investigación.

## 1.4. El problema de la eficiencia energética

La eficiencia energética se refiere a la práctica que tiene como objetivo la reducción del consumo de energía. En concreto en el sector de la industria, el precio en aumento de la energía y las regulaciones ambientales han dirigido a la industria a preocuparse

por la eficiencia energética como nunca antes. Un ejemplo es el programa ‘Think Blue. Factory’ de Volkswagen, el cual pretende reducir en 25 % el consumo de energía y gas, generación de desechos y emisión de CO<sub>2</sub> por cada vehículo manufacturado en 2018.

El problema de la eficiencia energética ha sido muy abordado en la literatura. Proyectos europeos del programa H2020 como NEED4B<sup>1</sup> y TRIBE<sup>2</sup> monitorizan diversos edificios, observan consumidores y modifican sus características físicas para que sean más eficientes. En [Tsanas and Xifara \[2012\]](#) utilizan técnicas de aprendizaje automático para estudiar el efecto de diversas características físicas, entre las que se incluyen la orientación del edificio o sus propiedades de acristalamiento, en las cargas de calentamiento y enfriamiento de diferentes edificios residenciales. El problema de predecir el consumo de energía futuro ha sido muy estudiado, y es interesante desde el punto de vista de la eficiencia energética porque permite contratar la energía necesaria, y no una cantidad mayor que se disiparía finalmente en el caso de no utilizarse. [Zhao and Magoulès \[2012\]](#) realizaron una revisión del estado del arte en este problema de predicción de la energía en edificios. También en [Shaikh et al. \[2014\]](#) se presenta una investigación exhaustiva sobre sistemas de control inteligentes en el estado del arte para la administración de energía y comodidad en los llamados edificios energéticamente inteligentes. Los autores recopilan un total de 121 trabajos relacionados. En esta línea de investigación, se ha demostrado que la presencia y comportamiento de los ocupantes en edificios tiene un gran impacto en el consumo de energía final. Consecuentemente, este aspecto ha sido foco de estudio y en [Nguyen and Aiello \[2013\]](#) se recoge una revisión para concluir con los principios y perspectivas en edificios energéticamente inteligentes basados en la actividad del usuario. Diversos modelos de redes Bayesianas han sido ampliamente utilizados en estas líneas de investigación, entre los que se incluyen los trabajos de [Hawarah et al. \[2010\]](#), [Lee and Cho \[2012\]](#), [Carbonari et al. \[2014\]](#), [Shoji et al. \[2015\]](#) y [Huang et al. \[2018\]](#).

En este proyecto pretendemos abordar el problema de la eficiencia energética desde una perspectiva innovadora y dentro del contexto de Industria 4.0. Dada una máquina industrial, queremos reducir su consumo energético mediante la acción de sincronizar las puestas en marcha y paradas de los elementos que la componen. Para ello, en primer lugar, es necesario conocer estos momentos de inicio y parada. Como no es posible situar un sensor en cada elemento activo de la máquina, ofreceremos un enfoque de acuerdo con los patrones generales de consumo de ésta. En el contexto de clasificación multi-dimensional, comenzaremos con ciertas variables predictoras que se miden a la entrada de la máquina, y trabajaremos en detectar (predecir) qué elementos

---

<sup>1</sup><http://need4b.eu/>

<sup>2</sup><http://tribe-h2020.eu/>

activos están en operación, de acuerdo a lo que se observa en la variable de consumo. Este proceso será posible gracias a que cada elemento activo tiene ligada una huella energética propia que le distingue (Figura 1.3).

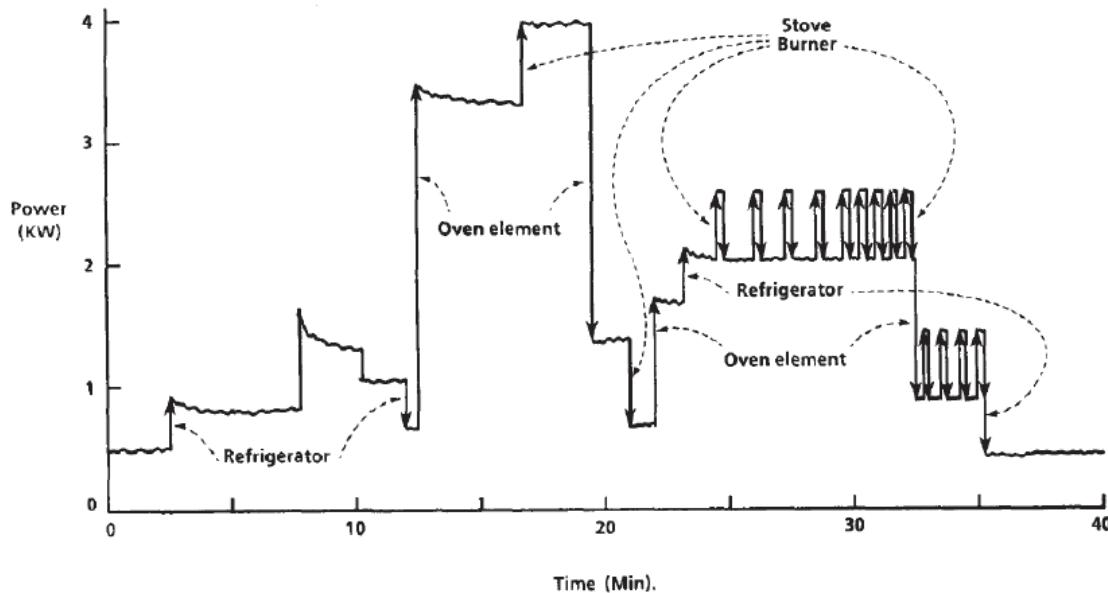


Figura 1.3: La potencia energética total de una vivienda mostrada a lo largo del tiempo refleja una serie de escalones en relación a los cambios de estado (encendido/apagado) de diferentes electrodomésticos. Fuente: [Hart \[1992\]](#).

Esta descomposición de la energía total en cargas individuales es ya conocida. En [Hart \[1992\]](#) se ofrece una aproximación muy simple para la detección de las cargas individuales, que consiste en medir inicialmente cuál es el cambio en el consumo total resultado de encender cada dispositivo de manera individual (su huella energética). Posteriormente, se monitoriza el consumo total y se buscan cambios similares a los anteriormente medidos, permitiendo de esta manera la detección de encendidos y apagados del dispositivo correspondiente. Si bien, esta sería una situación ideal en la que el consumo total corresponde a la suma de las cargas individuales. En la práctica, las cargas pueden ser inductivas, capacitivas y resistivas, lo que origina que éstas se compensen parcialmente y el consumo total no sea la suma directa. Es decir, el cambio en el consumo al encender un dispositivo será diferente en base a qué otros dispositivos están también encendidos. Por esta razón, en este trabajo queremos explorar la aproximación comentada basada en técnicas de aprendizaje automático.

Además, en este trabajo queremos explorar no sólo la detección del encendido y apagado de los diferentes elementos activos de la máquina, si no también sus diferentes estados de consumo, i.e., distinguir si el elemento se encuentra en máxima ejecución, está a una potencia intermedia, o solamente está en mínimo rendimiento. El trabajo

de Hart [1992] permite únicamente detectar estos cambios de estado ON/OFF, por lo que en este trabajo también queremos explorar la detección de los cambios de estado de consumo de los diferentes elementos activos de una máquina industrial.

## 1.5. Contribuciones

Este trabajo se centra en la investigación de nuevos clasificadores Bayesianos multi-dimensionales, además de su aplicación al problema de eficiencia energética en la Industria 4.0. En particular, hemos realizado las siguientes contribuciones:

- **Una revisión completa del estado del arte en clasificadores multi-dimensionales de redes Bayesianas.** El estudio de estos clasificadores Bayesianos especialmente diseñados para resolver problemas de clasificación multi-dimensional se establece como el pilar de investigación base de este trabajo. Se estudian todos los aspectos encontrados en la literatura en relación a estos clasificadores, entre los que se encuentran su definición, su complejidad de aprendizaje e inferencia, todos los métodos de aprendizaje de estos modelos y sus aplicaciones en dominios reales. También de interés, se revisa el conjunto de medidas de evaluación de rendimiento apropiadas para evaluar clasificadores multi-dimensionales. Toda esta revisión del estado del arte, además de incluir algunas aportaciones propias, se recoge en el **Capítulo 2** de este documento.
- **MBCTree:** un nuevo clasificador multi-dimensional híbrido de árboles de clasificación y clasificadores de redes Bayesianas multi-dimensionales. Hasta el mejor de nuestro conocimiento, este modelo es el primer híbrido propuesto en el contexto de clasificación multi-dimensional. También presentamos un método *wrapper* para aprender este nuevo clasificador desde un conjunto de datos. Finalmente, realizamos un estudio experimental llevado a cabo sobre conjuntos de datos sintéticos sobre el que se observan resultados favorables en términos de precisión predictiva. Este nuevo modelo creado y todas estas aportaciones comentadas se exponen en el **Capítulo 3** del documento.
- **Aplicación de técnicas de aprendizaje automático al problema de eficiencia energética en Industria 4.0.** Esta aplicación surge de la colaboración con la empresa Etxe-Tar S.A.<sup>3</sup> mediante iniciativa suya para desarrollar un método que se incorpore a su realidad empresarial. La empresa se encarga de proporcionarnos los datos de energía consumida en una de sus máquinas de mecanizado de una planta industrial, con los cuales nosotros somos

---

<sup>3</sup><http://www.etxe-tar.com/>

capaces de construir de manera supervisada un modelo multi-dimensional y así poder predecir qué elementos activos de la máquina están encendidos. En una segunda aproximación al problema también tenemos en cuenta los diferentes grados de consumo de los elementos activos y no solamente su estado de encendido o apagado. Todo lo relacionado con esta aplicación se recoge en el **Capítulo 4**.

- **Una revisión completa del estado del arte en la clasificación de flujos de datos multi-dimensionales con cambio de concepto.** En un trabajo futuro se pretende resolver el problema de eficiencia energética en tiempo real, de tal manera que recibamos los datos continuamente a diferencia de en un contexto estacionario. Estos flujos de datos se caracterizan por su potencial cambio de concepto a lo largo del tiempo, ya que las máquinas del proceso industrial se pueden deteriorar con el uso y consecuentemente afectar a la huella energética que muestran. Motivados por ello, en el **Capítulo 5** se ha querido realizar un estudio del estado del arte en la clasificación de flujos de datos multi-dimensionales en los que la relación entre las variables predictoras y la variable clase puede cambiar a lo largo del tiempo. De esta manera, la investigación llevada a cabo nos permitirá desarrollar un método futuro para la resolución del problema de eficiencia energética que se incorpore a una línea de producción en tiempo real.

## 1.6. Difusión científica

La investigación desarrollada en este trabajo pretende ser difundida en una serie de publicaciones científicas. En primer lugar, se está trabajando en un artículo en relación al nuevo clasificador multi-dimensional desarrollado que se detalla en el Capítulo 3, el cual se quiere publicar en la conferencia internacional IDEAL 2018<sup>4</sup>.

Por otra parte, se está trabajando en otra publicación para la revista científica de revisión por pares ACM Computing Surveys, con un JCR en el primer cuartil Q1, en relación a la revisión del estado del arte sobre los clasificadores multi-dimensionales de redes Bayesianas expuesta en el Capítulo 2 de este documento.

---

<sup>4</sup><https://aida.ii.uam.es/ideal2018/>



# Capítulo 2

## Clasificadores multi-dimensionales de redes Bayesianas

### 2.1. Fundamentos

#### 2.1.1. Definición

Una red Bayesiana [Pearl, 1988, Koller and Friedman, 2009] sobre un conjunto finito de variables discretas aleatorias  $\{Z_1, \dots, Z_n\}$ ,  $n \geq 1$ , es un par  $\mathcal{B} = (G, \Theta)$ .  $G = (V, A)$  es un grafo acíclico dirigido (DAG, del inglés *directed acyclic graph*) cuyos vértices  $V$  corresponden a las variables  $Z_i$  y cuyos arcos  $A$  representan dependencias probabilísticas directas entre los vértices.  $\Theta$  es un vector de parámetros tal que  $\theta_{z_i|\mathbf{pa}(z_i)} = p(z_i|\mathbf{pa}(z_i))$  define la probabilidad condicionada de cada posible valor  $z_i$  de  $Z_i$  dado un valor de vector  $\mathbf{pa}(z_i)$  de los padres de  $Z_i$  en  $G$ .  $\mathcal{B}$  representa una distribución de probabilidad conjunta  $p_{\mathcal{B}}$  sobre las variables factorizada de acuerdo a la estructura  $G$ :

$$p_{\mathcal{B}}(z_1, \dots, z_n) = \prod_{i=1}^n p(z_i|\mathbf{pa}(z_i)).$$

Los clasificadores de redes Bayesianas [Bielza and Larrañaga, 2014] son redes Bayesianas de topología restringida hechas a medida para resolver problemas de clasificación en los que instancias descritas por una serie de variables predictoras deben clasificarse en una única clase de un conjunto predefinido de varias clases diferentes. El conjunto finito de vértices  $V$  de un clasificador de redes Bayesianas se divide en un conjunto  $V_X = \{X_1, \dots, X_m\}$ ,  $m \geq 1$ , de variables predictoras y un conjunto unitario  $V_C = \{C\}$  que se corresponde con la variable clase (i.e.,  $n = m + 1$ ).

Un clasificador multi-dimensional de redes Bayesianas (MBC, del inglés *multi-dimensional Bayesian network classifier*) es una red Bayesiana especialmente diseñada para resolver problemas de clasificación multi-dimensional. El grafo  $G = (V, A)$  de un MBC tiene también el conjunto de vértices  $V$  dividido en dos conjuntos

$V_C = \{C_1, \dots, C_d\}$ ,  $d \geq 1$ , de variables clase y  $V_X = \{X_1, \dots, X_m\}$ ,  $m \geq 1$ , de variables predictoras (i.e.,  $n = m+d$ ). Los clasificadores de redes Bayesianas son así un escenario particular ( $d = 1$ ) de los MBCs. El grafo también tiene una topología restringida tal que el conjunto de arcos  $A$  se divide en tres subconjuntos  $A_C$ ,  $A_X$  y  $A_{CX}$ . La primera vez que los MBCs se propusieron por [van der Gaag and de Waal \[2006\]](#), los tres subconjuntos de arcos tenían las siguientes propiedades:

1. El subconjunto  $A_C \subseteq V_C \times V_C$  se compone de los arcos entre las variables clase, obteniendo un subgrafo  $G_C = (V_C, A_C)$ , *subgrafo clase*, de  $G$  inducido por  $V_C$ ;
2. El subconjunto  $A_X \subseteq V_X \times V_X$  se compone de los arcos entre las variables predictoras, obteniendo un subgrafo  $G_X = (V_X, A_X)$ , *subgrafo predictor*, de  $G$  inducido por  $V_X$ ;
3. El subconjunto  $A_{CX} \subseteq V_C \times V_X$  se compone de los arcos que van desde las variables clase hasta las variables predictoras, obteniendo un subgrafo  $G_{CX} = (V, A_{CX})$ , *subgrafo selección de predictoras*, de  $G$  inducido por  $V$ , tal que para cada  $X_i \in V_X$ , existe un  $C_j \in V_C$  con el arco  $(C_j, X_i) \in A_{CX}$  y para cada  $C_j \in V_C$ , existe un  $X_i \in V_X$  con el arco  $(C_j, X_i) \in A_{CX}$ .

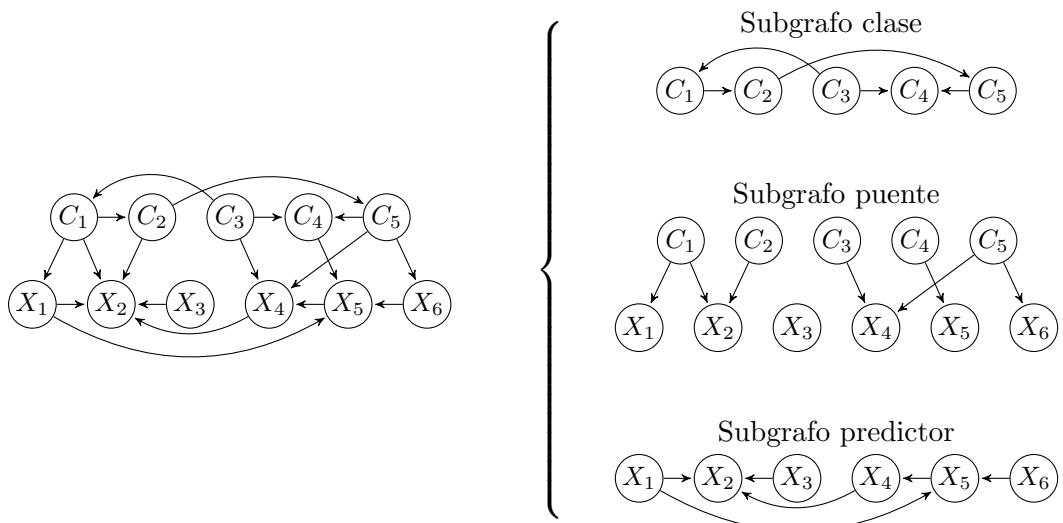


Figura 2.1: Ejemplo de la estructura de un MBC con sus tres subgrafos.

Los MBCs se extendieron posteriormente por [Bielza et al. \[2011\]](#), tal que se eliminaron las dos condiciones del subconjunto de arcos  $A_{CX}$  y el subgrafo resultante fue renombrado con otro término:

3. El subconjunto  $A_{CX} \subseteq V_C \times V_X$  se compone de los arcos que van desde las variables clase hasta las variables predictoras, obteniendo un subgrafo  $G_{CX} = (V, A_{CX})$ , *subgrafo puente*, de  $G$  inducido por  $V$ .

Esta última definición ha sido mayoritariamente adoptada en la literatura. La Figura 2.1 muestra un ejemplo de la estructura de un MBC y sus tres subgrafos. Se puede observar que la definición inicial de [van der Gaag and de Waal \[2006\]](#) no reconoce esta estructura como un MBC, dado que para  $X_3 \in V_X$ , no existe un  $C_j \in V_C$  con el arco  $(C_j, X_3) \in A_{CX}$ . La extensión de [Bielza et al. \[2011\]](#) se puede ver como una definición más general.

### 2.1.2. Familias de MBCs

Como detallaron [van der Gaag and de Waal \[2006\]](#), se pueden distinguir diferentes familias de MBCs en base a sus estructuras gráficas. Más tarde, [Bielza et al. \[2011\]](#) propusieron una notación que permitía una categorización completa de estos clasificadores. En general, los subgrafos clase y predictor pueden ser: vacíos, árboles, bosques de árboles, poliárboles, y de manera más general, DAGs. Las diferentes familias de MBCs se denotan como *estructura del subgrafo clase-estructura del subgrafo predictor*, donde las posibles estructuras son las cinco comentadas. Por esta razón, si ambos subgrafos clase y predictor son árboles, entonces la familia es un MBC árbol-árbol. La notación de completamente se introdujo por [van der Gaag and de Waal \[2006\]](#) cuando ambos subgrafos clase y predictor son el mismo tipo de estructura gráfica. Por ejemplo, un MBC árbol-árbol también se puede denotar como MBC completamente árbol. En la Figura 2.2 se exponen ejemplos de familias de MBCs.

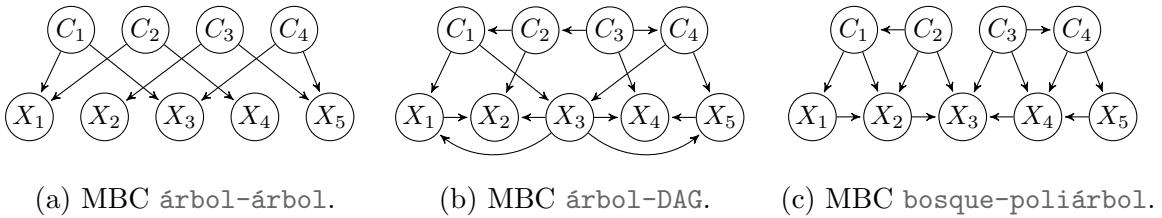


Figura 2.2: Ejemplos de estructuras gráficas en relación a diferentes familias de MBCs.

Se observa que clasificadores Bayesianos bien conocidos como el Bayes ingenuo (NB, del inglés *naive Bayes*) [[Minsky, 1961](#)], Bayes ingenuo selectivo [[Langley and Sage, 1999](#)], Bayes ingenuo aumentado en árbol [[Friedman et al., 1997](#)] (TAN, del inglés *tree-augmented naive Bayes*), Bayes ingenuo aumentado en árbol selectivo [[Blanco et al., 2005](#)] o los clasificadores Bayesianos  $k$ -dependientes ( $k$ -DB, del inglés  *$k$ -dependence Bayesian classifiers*) [[Sahami, 1996](#)] son casos especiales de MBCs donde

$d = 1$ . En la literatura se han extendido algunos de estos conocidos clasificadores al contexto multi-dimensional: NB y TAN multi-dimensionales [van der Gaag and de Waal, 2006] y  $k$ -DB multi-dimensional [Rodríguez and Lozano, 2008].

### 2.1.3. MBCs descomponibles

Bielza et al. [2011] definieron los MBCs clase-puente descomponibles (CB-MBCs, del inglés *class-bridge decomposable* MBCs) tal que:

1.  $G_C \cup G_{CX}$  se puede descomponer en  $G_C \cup G_{CX} = \bigcup_{i=1}^r (G_{C_{Comp_i}} \cup G_{(CX)_{Comp_i}})$ ,  $r \in \{2, \dots, d\}$ , donde  $G_{C_{Comp_i}} \cup G_{(CX)_{Comp_i}}$ , con  $i \in \{1, \dots, r\}$ , son sus  $r$  componentes conexos maximales<sup>1</sup>, y
2.  $\mathbf{Ch}(V_{C_{Comp_i}}) \cap \mathbf{Ch}(V_{C_{Comp_j}}) = \emptyset$ , con  $i, j \in \{1, \dots, r\}$  e  $i \neq j$ , donde  $\mathbf{Ch}(V_{C_{Comp_i}})$  denota los hijos de todas las variables en  $V_{C_{Comp_i}}$ , el subconjunto de variables clase en  $G_{C_{Comp_i}}$  (propiedad de no hijos compartidos).

En la Figura 2.3a se expone un ejemplo de un CB-MBC. Esta estructura tiene  $r = 2$  componentes conexos maximales, los cuales se muestran en la Figura 2.3b. El subgrafo a la izquierda de la línea vertical discontinua es  $G_{C_{Comp_1}} \cup G_{(CX)_{Comp_1}}$ , i.e., el primer componente conexo maximal, tal que  $V_{C_{Comp_1}} = \{C_1, C_2\}$  y  $\mathbf{Ch}(V_{C_{Comp_1}}) = \{X_1, X_2\}$ . De manera análoga, el subgrafo de la parte derecha es el segundo componente conexo maximal  $G_{C_{Comp_2}} \cup G_{(CX)_{Comp_2}}$ , tal que  $V_{C_{Comp_2}} = \{C_3, C_4, C_5\}$  y  $\mathbf{Ch}(V_{C_{Comp_2}}) = \{X_3, X_4, X_5, X_6\}$ . Se tiene que  $\mathbf{Ch}(\{C_1, C_2\}) \cap \mathbf{Ch}(\{C_3, C_4, C_5\}) = \emptyset$  tal como se requiere. Se observa que un CB-MBC siempre es un MBC bosque-estructura del subgrafo predictor, pero no vice versa (i.e., el MBC de la Figura 2.2c no es un CB-MBC porque  $\mathbf{Ch}(\{C_1, C_2\}) \cap \mathbf{Ch}(\{C_3, C_4\}) = \{X_3\} \neq \emptyset$ ).

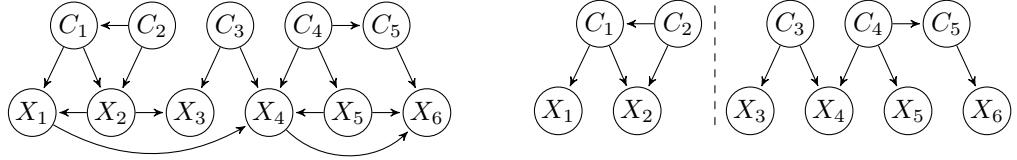
En la Sección 2.4 se incluyen los algoritmos propuestos en la literatura para aprender este tipo de estructuras descomponibles desde un conjunto de datos. Por otra parte, en la Sección 2.3.2 se detallan las ganancias obtenidas con los CB-MBCs durante el proceso de inferencia de nuevos ejemplos.

## 2.2. Medidas de evaluación de rendimiento

La evaluación de modelos en un contexto multi-dimensional necesita una aproximación especial dado que hay que tener en cuenta el rendimiento simultáneo sobre todas las variables clase. Varias medidas de evaluación de rendimiento se han

---

<sup>1</sup>Se dice que un grafo es conexo maximal si existe un camino entre cada par de vértices en su versión no dirigida.



(a) Estructura de un CB-MBC.

(b) Subgrafos conexos maximales.

Figura 2.3: Ejemplo de la estructura de un CB-MBC con sus dos subgrafos conexos maximales.

extendido al problema particular multi-etiqueta, pero sólo unas pocas extensiones para el problema más general de clasificación multi-dimensional se pueden encontrar en la literatura. En este trabajo contribuimos con la extensión de algunas medidas de evaluación de rendimiento multi-etiqueta al paradigma multi-dimensional.

### 2.2.1. Medidas multi-etiqueta en la literatura

Las medidas más frecuentes para clasificación multi-etiqueta se recogen en [Gibaja and Ventura \[2015\]](#) y se recopilan en la Tabla 2.1 siguiendo la taxonomía propuesta por [Tsoumakas et al. \[2009\]](#), la cual diferencia entre medidas para evaluar biparticiones y medidas para evaluar rankings. A continuación se resumen estas medidas multi-etiqueta haciendo especial interés en aquellas que tienen una extensión multi-dimensional.

#### Medidas para evaluar biparticiones

Las medidas para evaluar biparticiones se pueden clasificar en dos grupos: las basadas en etiquetas (*label-based*) y las basadas en ejemplos (*example-based*). Las primeras se calculan por cada etiqueta y después se promedian sobre todas ellas de dos maneras posibles, *macro* y *micro*. Las últimas se calculan por cada ejemplo de prueba y después se promedian a lo largo de todo el conjunto de prueba.

- *LABEL-BASED*. Para cada etiqueta, se calcula una medida uni-etiqueta  $B$  basada en el número de verdaderos positivos ( $vp$ ), verdaderos negativos ( $vn$ ), falsos positivos ( $fp$ ) y falsos negativos ( $fn$ ). Comúnmente,  $B$  es el  $accuracy = \frac{vp+vn}{vp+vp+tn+fn}$ ,  $precision = \frac{vp}{vp+fp}$ ,  $recall = \frac{vp}{vp+fn}$  o  $F_\beta = (1 + \beta) \frac{precision \cdot recall}{\beta \cdot precision + recall}$ . Se obtiene una matriz de confusión de dimensión  $2 \times 2$  por cada etiqueta, por lo que es necesario calcular un valor promedio.
  - La aproximación *macro* calcula una medida por cada etiqueta y después se promedian los valores obtenidos:

$$B_{macro} = \frac{1}{d} \sum_{j=1}^d B(vp_j, fp_j, vn_j, fn_j). \quad (2.1)$$

- La aproximación *micro* considera predicciones de todas las instancias juntas (agregando los valores de todas las matrices de confusión) y después se calcula la medida sobre todas las etiquetas:

$$B_{micro} = B \left( \sum_{j=1}^d tp_j, \sum_{j=1}^d fp_j, \sum_{j=1}^d tn_j, \sum_{j=1}^d fn_j \right). \quad (2.2)$$

Como detallaron [Gibaja and Ventura \[2015\]](#), estos dos tipos de promedio son informativos y no existe un acuerdo general sobre el uso de una aproximación *macro* o *micro*. Las puntuaciones promediadas con un enfoque *macro* otorgan un peso igual a cada etiqueta, independientemente de su frecuencia. Esto conduce a una mayor influencia del rendimiento de etiquetas raras. Sin embargo, las medidas *micro* dan el mismo peso a cada ejemplo y tienden a estar dominadas por el rendimiento en las etiquetas más comunes [\[Yang, 1999, Yang and Liu, 1999\]](#).

- *EXAMPLE-BASED*. El *0/1 subset accuracy* [\[Zhu et al., 2005\]](#), también llamado *precisión de clasificación* o *ratio de coincidencia exacta*, calcula la fracción de ejemplos correctamente clasificados, i.e., aquellos cuyo conjunto de etiquetas predichas es exactamente igual a su correspondiente conjunto de etiquetas verdaderas. Esta medida se puede ver como una extensión del *accuracy* tradicional al problema multi-etiqueta. Es una medida de evaluación muy estricta, especialmente cuando el tamaño del espacio de etiquetas,  $d$ , es grande, dado que una predicción completamente incorrecta y una predicción parcialmente correcta son considerados por igual un error de clasificación.

*Hamming loss* [\[Schapire and Singer, 1999\]](#) evalúa la fracción de parejas ejemplo-etiqueta mal clasificadas. Esta medida tiene en cuenta ambos errores de predicción (i.e., se predicen etiquetas irrelevantes) y errores de omisión (i.e., no se predicen etiquetas relevantes).

Por último, son comunes las medidas adoptadas desde el área de recuperación de información (RI) propuestas en [Godbole and Sarawagi \[2004\]](#). Este conjunto de medidas son específicas del paradigma multi-etiqueta y no tienen extensión al problema multi-dimensional, ya que miden el rendimiento de la recuperación de las etiquetas, las cuales pueden estar presentes o no. Si bien, en el dominio multi-dimensional una variable clase siempre va a estar presente, sea con un valor u otro. La única extensión posible sería si cada variable clase tuviese un valor semántico de “no presente” en relación al valor negativo de una etiqueta o variable binaria. En tal caso, el funcionamiento de estas medidas sobre un problema multi-dimensional sería igual que para un contexto multi-etiqueta.

## Medidas para evaluar rankings

Existen métodos como el propuesto en [Hüllermeier et al. \[2008\]](#) cuya salida es un ranking de relevancia entre las etiquetas del problema en cuestión. Este ranking indica que si una etiqueta cualquiera se debería asignar a una instancia, entonces todas aquellas que aparecen en posiciones anteriores en el ranking también se deberían asignar. Varias medidas se han propuesto en el contexto multi-etiqueta que comparan el ranking predicho para una instancia con su conjunto correspondiente de etiquetas verdaderas. Por ejemplo, en [Schapire and Singer \[2000\]](#) se propone la medida *one-error* que evalúa la fracción de ejemplos cuya etiqueta en el top del ranking no está en el conjunto de etiquetas relevantes verdaderas, y la medida *coverage*, la cual evalúa cuántas posiciones son necesarias, en media, para bajar en el ranking de etiquetas predicho de un ejemplo tal que se cubran todas sus etiquetas verdaderas.

De nuevo, este conjunto de medidas son específicas del paradigma multi-etiqueta y no tienen extensión al problema multi-dimensional. La razón se debe a que en un contexto multi-dimensional no tiene ningún sentido realizar un ranking entre los valores de las diferentes variables clase, ya que cada una va a tener siempre un valor asignado y cualquier ordenación de relevancia entre todos los valores carece de sentido.

### 2.2.2. Medidas multi-dimensionales en la literatura

#### Medidas de *accuracy*

[Bielza et al. \[2011\]](#) propusieron las siguientes medidas de evaluación de rendimiento que extienden aquellas en el dominio multi-etiqueta:

- El *global* o *joint accuracy* sobre la variable clase  $d$ -dimensional, la cual extiende la medida multi-etiqueta *0/1 subset accuracy* [[Zhu et al., 2005](#)] al calcular la fracción de ejemplos correctamente clasificados, i.e., aquellos cuyos valores clase predichos son los mismos que sus correspondientes valores verdaderos. Es una medida de evaluación muy estricta, especialmente cuando el tamaño del espacio de clases,  $|I|$ , es grande. Sea  $\mathbf{c}'_i$  la predicción  $d$ -dimensional para el ejemplo  $i$  en el conjunto de datos de prueba de  $N$  ejemplos,  $\mathbf{c}_i$  su correspondiente valor verdadero y  $\delta(\mathbf{c}'_i, \mathbf{c}_i) = 1$  si  $\mathbf{c}'_i = \mathbf{c}_i$  y 0 en caso contrario, entonces el *global accuracy* se define como:

$$Acc = \frac{1}{N} \sum_{i=1}^N \delta(\mathbf{c}'_i, \mathbf{c}_i). \quad (2.3)$$

- El *mean* o *average accuracy* sobre las  $d$  variables clase, el cual evalúa la fracción de parejas ejemplo-clase correctamente clasificadas. Sea  $c'_{ij}$  el valor clase predicho para la variable clase  $C_j$  y el ejemplo  $i$  del conjunto de datos de prueba,  $c_{ij}$  su

correspondiente valor verdadero y  $\delta(c'_{ij}, c_{ij}) = 1$  si  $c'_{ij} = c_{ij}$  y 0 en caso contrario. Entonces, el *mean accuracy* se define de la siguiente manera:

$$\overline{Acc}_d = \frac{1}{d} \sum_{j=1}^d Acc_j = \frac{1}{d} \sum_{j=1}^d \frac{1}{N} \sum_{i=1}^N \delta(c'_{ij}, c_{ij}). \quad (2.4)$$

Esta medida es la complementaria de la multi-etiqueta *Hamming loss* [Schapire and Singer, 1999], i.e., *mean accuracy*+*Hamming loss* = 1, pero extendida al paradigma multi-dimensional.

- Una idea similar fue aplicada a los CB-MBCs por medio del *mean accuracy* sobre los  $r$  componentes conexos maximales:

$$\overline{Acc}_r = \frac{1}{r} \sum_{j=1}^r \frac{1}{N} \sum_{i=1}^N \delta(\mathbf{c}_i^{\downarrow V_{C_{Comp_j}}}, \mathbf{c}_i), \quad (2.5)$$

donde  $\mathbf{c}_i^{\downarrow V_{C_{Comp_j}}}$  representa la proyección del vector  $\mathbf{c}_i$  en las coordenadas de  $V_{C_{Comp_j}}$ . Se cumple que  $Acc \leq \overline{Acc}_r \leq \overline{Acc}_d$  dado que es más estricto contar predicciones correctas sobre un vector de componentes en su conjunto que de manera individual sobre sus componentes.

## Medidas de calibración probabilísticas

Las arriba mencionadas medidas de *accuracy*, así como sus versiones uni-dimensional y multi-etiqueta, evalúan el rendimiento de un modelo solamente considerando la clasificación final que realiza éste. Los modelos probabilísticos, como son las redes Bayesianas, nos ofrecen más información, i.e., la probabilidad estimada *a posteriori* de cada valor clase, que la propia clasificación final como tal, la cual es de hecho el valor clase que maximiza esta probabilidad estimada (bajo una función de pérdida estándar 0/1). El *Brier score* [Brier, 1950] mide la calibración de modelos probabilísticos al tener en cuenta las probabilidades estimadas *a posteriori*, tal que aquellos clasificadores que estén casi seguros al hacer predicciones (correctas) tendrán un valor más bajo del *Brier score*. En un problema uni-dimensional en el que una única variable clase  $C$  se clasifica como un valor de  $|\Omega_C|$  posibles, tal que  $c_k$  es su  $k$ -ésimo valor clase,  $c_i$  el valor verdadero para el ejemplo  $i$  descrito con las predictoras  $\mathbf{x}_i$  y el resto de símbolos definidos como antes, entonces el *Brier score* se define como:

$$Bs = \frac{1}{N} \sum_{i=1}^N \sum_{k=1}^{|\Omega_C|} [p(C = c_k | \mathbf{x}_i) - \delta(c_k, c_i)]^2.$$

El *Brier score* se generalizó por [Fernandes et al. \[2013\]](#) al problema de clasificación multi-dimensional en las tres variantes que se describen a continuación.

- El *global* o *joint Brier score*:

$$Bs = \frac{1}{N} \sum_{i=1}^N \sum_{g=1}^{|I|} [p(\mathbf{C} = \mathbf{c}_g | \mathbf{x}_i) - \delta(\mathbf{c}_g, \mathbf{c}_i)]^2, \quad (2.6)$$

donde  $\mathbf{C} = (C_1, \dots, C_d)$  es la variable clase  $d$ -dimensional,  $\mathbf{c}_g$  es la  $g$ -ésima configuración de  $I$  y  $\mathbf{c}_i$  es el valor verdadero de  $\mathbf{C}$  para  $\mathbf{x}_i$ .

- El *mean* o *average Brier score*:

$$\overline{Bs}_d = \frac{1}{d} \sum_{j=1}^d Bs_j = \frac{1}{d} \sum_{j=1}^d \frac{1}{N} \sum_{i=1}^N \sum_{k=1}^{|\Omega_{C_j}|} [p(C_j = c_{jk} | \mathbf{x}_i) - \delta(c_{jk}, c_{ij})]^2, \quad (2.7)$$

donde  $c_{jk}$  es el  $k$ -ésimo valor clase de la  $j$ -ésima variable clase,  $C_j$ , y  $c_{ij}$  es el valor verdadero de  $C_j$  para  $\mathbf{x}_i$ .

Se observa que ambas generalizaciones del *Brier score* también se encuentran en el rango  $[0, 2]$ , como sucede en la versión uni-dimensional. De igual manera, el modelo está más calibrado a menor valor del *Brier score*, tal que el valor cero se obtiene cuando el modelo predice el valor verdadero con total certeza y para todos los ejemplos del conjunto de prueba, y el valor dos se consigue cuando el modelo predice con total certeza una configuración que difiere del valor verdadero en todas las variables clase y para todos los ejemplos del conjunto de prueba.

Aunque ambas medidas generalizadas consideran la probabilidad estimada de cada clase, el *global Brier score* solamente recompensa la probabilidad estimada de la configuración de clases que coincide con el valor verdadero, y el *mean Brier score* recompensa las variables clase de manera separada. Por esta razón, los autores propusieron una tercera medida para premiar el número de clases que se clasifican correctamente de cada configuración de  $\mathbf{C}$ , tal que se obtenga un menor valor de esta medida cuando se asignen probabilidades más altas a configuraciones que estén cerca del valor verdadero.

- El *multi-dimensional calibrated Brier score*<sup>2</sup>:

---

<sup>2</sup>Es importante especificar que hemos modificado el término  $r_s = \sum_{j=1}^d |\Omega_{C_j}|$  en [Fernandes et al. \[2013\]](#) por  $d$  en el denominador de la ecuación para que la medida esté correctamente normalizada entre 0 y 1.

$$CBs = \frac{1}{Nd} \sum_{i=1}^N \sum_{g=1}^{|I|} p(\mathbf{C} = \mathbf{c}_g | \mathbf{x}_i) \sum_{j=1}^d (1 - \delta(c_{gj}, c_{ij})), \quad (2.8)$$

donde  $c_{gj}$  es valor clase de la variable clase  $C_j$  de la  $g$ -ésima configuración,  $\mathbf{c}_g$ .

### 2.2.3. Medidas multi-dimensionales extendidas

En esta sección realizamos dos contribuciones a las métricas de evaluación de rendimiento para modelos que resuelven problemas de clasificación multi-dimensional.

La primera contribución sigue la misma idea de extender el *mean accuracy* sobre los CB-MBCs (ec. (2.5)) pero aplicado al *Brier score*. Sea  $I_j = \times_{C_k \in V_{C_{Comp_j}}} \Omega_{C_k}$  el espacio de configuraciones de las variables clase del componente conexo maximal  $j$  y  $\mathbf{c}_{jg}$  la configuración  $g$ -ésima del espacio de configuraciones del propio componente  $j$ , entonces el *mean Brier score* sobre los  $r$  componentes conexos maximales es:

$$\overline{Bs}_r = \frac{1}{r} \sum_{j=1}^r \frac{1}{N} \sum_{i=1}^N \sum_{g=1}^{|I_j|} \left[ p\left(\mathbf{C}^{\downarrow V_{C_{Comp_j}}} = \mathbf{c}_{jg} | \mathbf{x}_i\right) - \delta\left(\mathbf{c}_{jg}, \mathbf{c}_i^{\downarrow V_{C_{Comp_j}}}\right) \right]^2. \quad (2.9)$$

Esta medida también se encuentra en el rango  $[0, 2]$ , y al igual que sucedía con las medidas de *accuracy*, se cumple que  $Bs \leq \overline{Bs}_r \leq \overline{Bs}_d$ .

La segunda contribución es la extensión de las medidas *label-based* del dominio multi-etiqueta al problema de clasificación multi-dimensional. A estas medidas las llamamos basadas en clase (*class-based*). Siguiendo la misma idea, se obtiene una matriz de confusión de dimensión  $|\Omega_{C_j}| \times |\Omega_{C_j}|$  para cada variable clase  $C_j$ .

- La aproximación *macro* calcula una medida para cada valor de las variables clase y después se promedian los resultados. Si una variable clase  $C_j$  es binaria, i.e.,  $|\Omega_{C_j}| = 2$ , solamente se calcula una medida para una de las dos clases, de tal manera que se evita redundancia (los verdaderos positivos de una clase serán los verdaderos negativos de la otra, y lo mismo sucede con los falsos positivos y falsos negativos). Sea  $vp_{jk}$  los verdaderos positivos del  $k$ -ésimo valor de la  $j$ -ésima variable clase,  $C_j$ , y de manera análoga para el resto de contadores, entonces:

$$B_{macro} = \frac{1}{d} \sum_{j=1}^d B_j, \quad \text{donde } B_j = \begin{cases} \frac{1}{|\Omega_{C_j}|} \sum_{k=1}^{|\Omega_{C_j}|} B(vp_{jk}, fp_{jk}, vn_{jk}, fn_{jk}), & \text{si } |\Omega_{C_j}| > 2 \\ B(vp_j, fp_j, vn_j, fn_j), & \text{si } |\Omega_{C_j}| = 2 \end{cases} \quad (2.10)$$

- La aproximación *micro* sigue la idea de agregar los valores de todas las matrices de confusión. Sin embargo, dentro de cada variable clase se realiza una normalización para que aquellas variables con muchos posibles valores no tengan una mayor influencia en el resultado final. De nuevo se evita la arriba comentada redundancia, por lo que la medida se define como:

$$B_{micro} = B \left( \sum_{j=1}^d VP_j, \sum_{j=1}^d FP_j, \sum_{j=1}^d VN_j, \sum_{j=1}^d FN_j \right), \text{ donde}$$

$$\{VP_j, FP_j, VN_j, FN_j\} = \begin{cases} \frac{1}{|\Omega_{C_j}|} \sum_{k=1}^{|\Omega_{C_j}|} \{vp_{jk}, fp_{jk}, vn_{jk}, fn_{jk}\}, & \text{si } |\Omega_{C_j}| > 2 \\ \{vp_j, fp_j, vn_j, fn_j\}, & \text{si } |\Omega_{C_j}| = 2 \end{cases} \quad (2.11)$$

Tabla 2.1: Equivalencia entre medidas de evaluación de rendimiento para problemas de clasificación multi-etiqueta y multi-dimensional.

	Multi-etiqueta	Multi-dimensional
Label-based	$B_{macro}$ (ec. (2.1)) $B_{micro}$ (ec. (2.2))	$B_{macro}$ (ec. (2.10)) $B_{micro}$ (ec. (2.11))
	0/1 subset accuracy [Zhu et al., 2005]	Global accuracy (ec. (2.3)) [Bielza et al., 2011]
	Hamming loss [Schapire and Singer, 1999]	Mean accuracy (ecs. (2.4), (2.5)) [Bielza et al., 2011]
Biparticiones	Recall	
Example-based	Precision Accuracy F1-score [Gödbole and Sarawagi, 2004]	<div style="display: flex; align-items: center;"> <span style="margin-right: 10px;">}</span> <span style="font-size: 2em;">RI</span> <span style="margin-left: 10px;">-</span> </div>
		Brier score (ecs. (2.6), (2.7), (2.8), (2.9)) [Fernandes et al., 2013]
	One-error [Schapire and Singer, 2000]	-
	Coverage [Schapire and Singer, 2000]	-
	Ranking loss [Schapire and Singer, 1999]	-
	IsError [Mencía and Fürnkranz, 2010]	-
Rankings	Average precision [Schapire and Singer, 2000]	-
	Margin loss [Mencía and Fürnkranz, 2010]	-
	Ranking error [Park and Fürnkranz, 2008]	-

## 2.3. Complejidad en los MBCs

### 2.3.1. Aprendizaje: cardinalidad del espacio de estructuras

Saber la cardinalidad del espacio de estructuras de MBCs nos puede ayudar a inferir la complejidad del problema de aprendizaje de estos modelos. Bielza et al. [2011] calcularon el número de todas las posibles estructuras de MBCs, puntuizando dos casos. El primero es la definición general de un MBC por Bielza et al. [2011], mientras que el segundo es la definición inicial de van der Gaag and de Waal [2006], la cual establece dos condiciones en el subgrafo puente de los MBCs: (a) para cada  $X_i \in V_X$ , existe un  $C_j \in V_C$  con el arco  $(C_j, X_i) \in A_{CX}$  y (b) para cada  $C_j \in V_C$ , existe un  $X_i \in V_X$  con el arco  $(C_j, X_i) \in A_{CX}$ .

1. El número de todas las posibles estructuras de MBCs con  $d$  variables clase y  $m$  variables predictoras,  $MBC(d, m)$ , es [Bielza et al., 2011, Teorema 6]:

$$MBC(d, m) = S(d) \cdot 2^{dm} \cdot S(m), \text{ donde}$$

$$S(n) = \sum_{i=1}^n (-1)^{i+1} \binom{n}{i} 2^{i(n-i)} S(n-i)$$

es la fórmula de Robinson [Robinson, 1973] que cuenta el número de posibles DAGs con  $n$  nodos, la cual se inicializa en  $S(0) = S(1) = 1$ . De esta manera,  $S(d)$  y  $S(m)$  cuentan las posibles estructuras de DAG para los subgrafos clase y predictor, respectivamente.  $2^{dm}$  es el número de posibles subgrafos puente.

2. El número de todas las posibles estructuras de MBCs con  $d$  variables clase y  $m$  variables predictoras,  $m \geq d$ , que satisfacen las condiciones (a) y (b),  $MBC^{ab}(d, m)$ , es [Bielza et al., 2011, Teoremas 7 y 8]:

$$MBC^{ab}(d, m) = S(d) \cdot BRS(d, m) \cdot S(m),$$

donde  $BRS(d, m)$  es el número de posibles subgrafos puente para los MBCs que satisfacen las dos condiciones requeridas,

$$BRS(d, m) = \sum_{k=m}^{dm} BRS(d, m, k),$$

calculado a partir de todos los subgrafos puente con  $k$  arcos,  $BRS(d, m, k)$ , con  $k$  mayor que  $m$  para que ninguna variable predictora se quede sin conectar, y hasta el máximo posible de arcos,  $dm$ , tal que:

$$BRS(d, m, k) = \binom{dm}{k} - \sum_{\substack{x \leq d, y \leq m \\ k \leq xy \leq dm-d}} \binom{d}{x} \binom{m}{y} BRS(x, y, k).$$

De todos los posibles subgrafos puente con  $k$  arcos,  $\binom{dm}{k}$ , se eliminan aquellos inválidos que no cumplen las dos condiciones, sabiendo que  $k = dm - d + 1$  es el número mínimo de arcos con los que un subgrafo puente siempre es válido. Esta recursión se inicializa en  $BRS(1, 1, 1) = BRS(1, 2, 2) = BRS(2, 1, 2) = 1$ .

Los autores desarrollaron que:

$$BRS(d, m) = 2^{dm} - \sum_{k=0}^{m-1} \binom{dm}{k} - \sum_{k=m}^{dm} \sum_{\substack{x \leq d, y \leq m \\ k \leq xy \leq dm-d}} \binom{d}{x} \binom{m}{y} BRS(x, y, k),$$

por lo que  $BRS(d, m) < 2^{dm}$ , para todos los valores de  $d \geq 1$  y  $m \geq 1$ , dado que los subgrafos puente que no satisfacen las condiciones requeridas se eliminan del número subgrafos totales,  $2^{dm}$ . De esta manera, se cumple que  $MBC^{\text{ab}}(d, m) < MBC(d, m)$ , para todos los  $d \geq 1$  y  $m \geq 1$ . La complejidad de la fórmula de Robinson se demostró ser súper exponencial, i.e.,  $O(S(n)) = n^{2^{O(n)}}$ , por lo que la complejidad del espacio de estructuras de los MBCs se calcula como [Bielza et al., 2011, Corolario 4]:

$$O(MBC(d, m)) = O(MBC^{\text{ab}}(d, m)) = 2^{dm} (\max\{d, m\})^{2^{O(\max\{d, m\})}}.$$

En este trabajo hemos querido extender el conocimiento de estas complejidades mediante el cálculo del número de estructuras CB-MBC con  $d$  variables clase,  $m$  variables predictoras y  $r$  componentes conexos maximales,  $CB\text{-}MBC^{\text{ab}}(d, m, r)$ , tal que se cumplen las dos condiciones en el subgrafo puente de la definición en [van der Gaag and de Waal \[2006\]](#), por lo que necesariamente  $r \leq \min(d, m)$ :

$$\begin{aligned} CB\text{-}MBC^{\text{ab}}(d, m, r) &= \\ &= \frac{\sum_{x=1}^{d-r+1} \binom{d}{x} \sum_{y=1}^{m-r+1} \binom{m}{y} CB\text{-}MBC^{\text{ab}}(x, y, 1) \cdot CB\text{-}MBC^{\text{ab}}(d-x, m-y, r-1)}{r!}. \end{aligned}$$

En esta recursión, elegimos para el primer componente conexo maximal todas las combinaciones de  $x$  variables clase e  $y$  predictoras hasta  $d - r + 1$  y  $m - r + 1$ , respectivamente, para que cada componente siguiente siga teniendo al menos una variable clase y predictora. Las combinaciones de escoger las  $x$  variables clase e  $y$  variables predictoras son  $\binom{d}{x}$  y  $\binom{m}{y}$ , respectivamente. Sobre ellas, calculamos el número de estructuras MBC que no sean descomponibles, es decir, que tengan un único componente conexo maximal. La recursión sigue mediante el cálculo del número de estructuras CB-MBC para los  $r - 1$  componentes conexos siguientes y sobre las  $d - x$  variables clase y  $m - y$  predictoras restantes. Es necesaria una división entre  $r!$  del número total de estructuras para eliminar los órdenes entre los componentes, ya que se van a calcular estructuras exactamente iguales pero distinguiendo cuál es el primer componente creado y sucesivamente hasta el último. La recursión se detiene mediante el cálculo del número de estructuras del último componente conexo. Para ello, hay que forzar a que un único componente conexo no sea propiamente descomponible:

$$CB\text{-}MBC^{\text{ab}}(d, m, 1) = MBC^{\text{ab}}(d, m) - \sum_{r=2}^{\min(d, m)} CB\text{-}MBC^{\text{ab}}(d, m, r),$$

lo que conseguimos mediante la eliminación de todas las estructuras descomponibles para todas las estructuras MBC posibles.

Para dar una mejor idea de la cardinalidad de los espacios de estructuras comentados, en la Tabla 2.2 se recogen diferentes cálculos sobre distintas configuraciones de variables clase y predictoras. Se observa claramente el crecimiento súper exponencial del número de estructuras a mayor número de variables, lo que deja evidente la necesidad de algoritmos de aprendizaje que se muevan de manera eficiente por este espacio. De igual forma, se aprecia que la definición de un MBC de [Bielza et al. \[2011\]](#) es más general y acepta un mayor número de estructuras válidas frente a la definición con las dos restricciones de [van der Gaag and de Waal \[2006\]](#), aunque esto no suponga gran diferencia y ambas se mantengan en el mismo orden de magnitud para las configuraciones mostradas. Si éstas se comparan con las posibles estructuras de DAG,  $S(d + m)$ , sí se observa una diferencia notable en el número de estructuras, justificada por la restricción de arcos desde variables predictoras a variables clase en los MBCs. Finalmente, el espacio de estructuras descomponibles CB-MBCs es varias órdenes de magnitud menor respecto a los MBCs más generales.

Tabla 2.2: Número de estructuras MBC para diferentes valores de variables clase  $d$ , variables predictoras  $m$  y componentes conexos maximales  $r$ .

$d$	$m$	$r$	$CB\text{-}MBC^{\text{ab}}(d, m, r)$	$MBC^{\text{ab}}(d, m)$	$MBC(d, m)$	$S(d + m)$
2	3	2	18	1.875	4.800	29.281
3	4	2	28.278	$2.96 \cdot 10^7$	$5.56 \cdot 10^7$	$1.14 \cdot 10^9$
		3	54			
		6	$4.07 \cdot 10^8$	$1.09 \cdot 10^{13}$	$2.48 \cdot 10^{13}$	$1.21 \cdot 10^{15}$
	6	3	$3.92 \cdot 10^4$			
		2	$2.96 \cdot 10^{11}$			
		3	$1.54 \cdot 10^7$	$2.24 \cdot 10^{16}$	$3.44 \cdot 10^{16}$	$4.18 \cdot 10^{18}$
4	4	4	$1.81 \cdot 10^3$			
		2	$4.07 \cdot 10^{21}$			
	9	3	$1.66 \cdot 10^{15}$	$2.52 \cdot 10^{28}$	$4.53 \cdot 10^{28}$	$1.87 \cdot 10^{31}$
		4	$8.32 \cdot 10^8$			

### 2.3.2. Inferencia: tratabilidad de las explicaciones más probables

En esta sección revisamos la literatura sobre el estudio de la complejidad del proceso de inferencia en los MBCs. La clasificación multi-dimensional implica tener una función de perdida  $\lambda(\mathbf{c}', \mathbf{c})$  que represente para cada pareja de vectores  $\mathbf{c}' , \mathbf{c} \in I$  el coste de clasificar un ejemplo como  $\mathbf{c}'$  cuando su valor verdadero es  $\mathbf{c}$ . En la literatura se ha estudiado el problema de clasificación multi-dimensional en los MBCs sobre dos funciones de pérdida. La primera es la función estándar 0-1 que asigna una unidad de pérdida a cualquier error, i.e., siempre que  $\mathbf{c}' \neq \mathbf{c}$ , y ninguna pérdida cuando el modelo realiza una clasificación correcta, i.e., cuando  $\mathbf{c}' = \mathbf{c}$ . Las segundas son las funciones de pérdida aditivas CB-decomponibles las cuales se ajustan a estructuras CB-MBCs.

#### Funciones de pérdida 0-1

Sea  $R(\mathbf{c}'|\mathbf{x}) = \sum_{g=1}^{|I|} \lambda(\mathbf{c}', \mathbf{c}_g) p(\mathbf{c}_g|\mathbf{x})$  la pérdida esperada o riesgo condicional, donde  $\mathbf{x} = (x_1, \dots, x_m)$  es un vector de valores de predictoras y  $p(\mathbf{c}_g|\mathbf{x})$  es la probabilidad conjunta posterior asignada por el modelo a la configuración de clases  $\mathbf{c}_g$  dada la observación  $\mathbf{x}$ . Entonces, bajo una función de pérdida 0-1, la regla de decisión de Bayes que minimiza la pérdida esperada  $R(\mathbf{c}'|\mathbf{x})$  es equivalente a seleccionar la configuración  $\mathbf{c}_g$  que maximice la probabilidad posterior  $p(\mathbf{c}_g|\mathbf{x})$  [Bielza et al., 2011, Teorema 1]:

$$\min_{\mathbf{c}'} R(\mathbf{c}'|\mathbf{x}) \Leftrightarrow \max_{\mathbf{c}_g} p(\mathbf{c}_g|\mathbf{x}).$$

De esta manera, la clasificación multi-dimensional bajo una función de pérdida

0-1 y una observación completa de todas las variables predictoras es equivalente a calcular la explicación más probable o abducción total (MPE, del inglés *most probable explanation*). Este problema es un tipo de máximo a posteriori (MAP), en el cual no hay necesidad de observar todas las variables predictoras. La investigación existente aborda la complejidad de la clasificación multi-dimensional en MBCs como la complejidad de calcular el MPE, i.e., se conocen todos los valores de las variables predictoras, dado que en caso contrario se obtendría la explicación más probable de las clases y las predictoras faltantes, lo cual puede diferir del MAP de únicamente las variables clase.

Se ha demostrado que calcular la MPE es un problema NP-duro para redes Bayesianas [Shimony, 1994], así como aproximarla con una precisión deseada [Abdelbar and Hedetniemi, 1998]. En la literatura se pueden encontrar algoritmos que resuelven de manera exacta este problema mediante *junction trees* [Dawid, 1992, Dechter, 1999], eliminación de variables [Li and D'Ambrosio, 1993] o búsqueda por ramificación y poda [Kask and Dechter, 2001, Marinescu and Dechter, 2009], y algoritmos aproximados basados en algoritmos genéticos [Gelsema, 1995, Rojas-Guzman and Kramer, 1993], algoritmos de búsqueda local estocástica [Kask and Dechter, 1999, Hutter et al., 2005], eliminación de variables [Dechter and Rish, 1997], búsqueda voraz el primero el mejor [Shimony and Charniak, 1990] y programación lineal [Santos, 2014].

Para obtener la MPE es necesario así calcular las probabilidades posteriores de todas las configuraciones en *I*. Bielza et al. [2011] se motivan por la similaridad existente entre las probabilidades de aquellas configuraciones con los mismos valores de clase excepto uno para enumerar este espacio de configuraciones *I*. Para ello, los autores proponen una extensión de la adaptación del código Gray presentada en Guan [1998] que permite enumerar configuraciones de variables clase con distinto número de posibles valores, tal que cada par de configuraciones adyacentes difiere en un solo componente y la diferencia es 1 ó -1. Los autores ofrecen una cota de las ganancias máximas obtenidas en el número de factores necesarios para calcular las probabilidades posteriores y por tanto la MPE mediante esta manera especial de moverse por el espacio de configuraciones frente a una aproximación por fuerza bruta [Bielza et al., 2011, Teorema 2].

De manera adicional, los mismos autores explotan la estructura especial de los CB-MBCs y demuestran que el problema de maximización para calcular el MPE se puede transformar en *r* sub-problemas independientes de maximización sobre espacios dimensionales más pequeños, i.e., sobre cada espacio de configuraciones *I<sub>j</sub>* de las variables clase del componente conexo maximal *j*:

$$\max_{\mathbf{c}_g} p(\mathbf{c}_g | \mathbf{x}) \propto \prod_{j=1}^r \max_{\mathbf{c}_g \in I_j} \phi_j^{\mathbf{x}} \left( \mathbf{c}_g^{\downarrow V_{C_{Comp_j}}} \right), \text{ donde}$$

$$\phi_j^{\mathbf{x}} \left( \mathbf{c}_g^{\downarrow V_{C_{Comp_j}}} \right) = \prod_{C_k \in V_{C_{Comp_j}}} p(c_{gk} | \mathbf{pa}(c_{gk})) \prod_{X_i \in \mathbf{Ch}(V_{C_{Comp_j}})} p(x_i | \mathbf{pa}_{V_C}(x_i), \mathbf{pa}_{V_X}(x_i)).$$

Los autores demostraron que se cumple:

$$\phi_j^{\mathbf{x}} \left( \mathbf{c}_g^{\downarrow V_{C_{Comp_j}}} \right) \propto p \left( \mathbf{C}^{\downarrow V_{C_{Comp_j}}} = \mathbf{c}_g^{\downarrow V_{C_{Comp_j}}} | \mathbf{x} \right).$$

La misma idea de enumerar todas las configuraciones de clases con una extensión del código Gray se aplica en [Bielza et al. \[2011\]](#), Teorema 4] sobre cada componente maximal conexo de un CB-MBC, conduciendo a unas ganancias aún mayores en el número de factores necesarios para calcular las probabilidades posteriores de las distintas configuraciones de clases en el cómputo de la MPE.

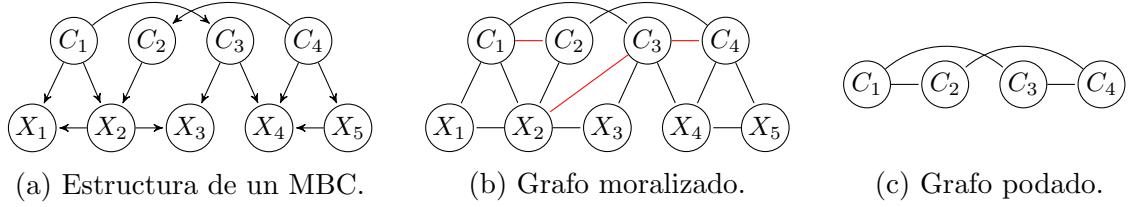


Figura 2.4: Ejemplo de moralización y poda de la estructura de un MBC.

Aunque el problema de calcular la MPE en una red Bayesiana  $\mathcal{B}$  es generalmente NP-duro, éste se puede calcular en tiempo polinómico en  $\mathcal{B}$  si la anchura de árbol,  $treewidth(G)$ , de su estructura  $G$  está limitada [[Sy, 1992](#)]. La anchura de árbol de un grafo dirigido  $G$  es la anchura de su moralización, i.e., del grafo no dirigido resultado de conectar los padres de cada variable y posteriormente eliminar las direcciones de los arcos (Figura 2.4b). Si se añaden las restricciones de estructura de los MBCs, [de Waal and van der Gaag \[2007\]](#) demostraron en su Teorema 1 que:

$$treewidth(G) \leq treewidth(G_X) + d, \quad (2.12)$$

donde  $G_X$  es el subgrafo predictor del MBC y  $d$  el número de variables clase. Esto significa que se puede realizar clasificación multi-dimensional sobre un MBC en tiempo polinómico si la suma de la anchura del subgrafo predictor y el número de variables clase está limitada, y que la conexión del subgrafo clase es irrelevante para la tratabilidad de la clasificación. [Kwisthout \[2011\]](#) aplicó esta misma idea a las estructuras CB-MBCs:

$$\text{treewidth}(G) \leq \text{treewidth}(G_X) + |d_{max}|,$$

donde  $|d_{max}|$  es el número de variables clase del componente conexo maximal con el máximo número de variables clase. De esta manera, la MPE se puede calcular en tiempo polinómico si la anchura de árbol de  $\text{treewidth}(G_X)$  y el número de variables de cada componente de  $G$  están limitados.

[Pastink and van der Gaag \[2015\]](#) se centraron en MBCs con un subgrafo predictor vacío, acotando su estructura de la siguiente manera:

$$\text{treewidth}(G_{\bar{F}}) < \text{treewidth}(G'), \quad (2.13)$$

donde  $G_{\bar{F}}$  es la estructura de un MBC con su subgrafo predictor vacío y  $G'$  es el denominado grafo podado, resultado de moralizar  $G_{\bar{F}}$  y después eliminar las variables predictoras. En la Figura 2.4 se muestra este proceso en un MBC general.

Finalmente, [Benjumeda et al. \[2018\]](#) se motivan por la dependencia entre una consulta cualquiera sobre una red Bayesiana y su complejidad de inferencia, i.e., los parámetros de la red se pueden actualizar con los valores de las variables evidencia de la consulta antes de realizar la inferencia. De esta manera, extienden el teorema anterior aplicado sobre MBCs con un subgrafo predictor vacío a cualquier MBC DAG-DAG, tal que se puede calcular la MPE en tiempo polinómico si la anchura de árbol de su grafo podado y el número de padres de cada variable evidencia están limitados. Los autores sostienen que aunque calcular la anchura de árbol del grafo podado  $G'$  es menos costoso computacionalmente que calcular la anchura de la estructura completa  $G$ , este primer problema sigue siendo NP-duro de calcular exactamente [[Arnborg et al., 1987](#)]. Por ello, observan que  $\text{treewidth}(G') \leq d$ , concluyendo así que el cálculo de la MPE es polinómico si el número de variables clase  $d$  y el número de padres de cada variable evidencia están limitados. Los autores extienden el mismo razonamiento sobre estructuras CB-MBCs en base a la limitación del número de variables clase de cada componente conexo maximal.

### Funciones de pérdida aditivas CB-decomponibles

Sea  $\lambda(\mathbf{c}', \mathbf{c})$  una función de pérdida, [Bielza et al. \[2011\]](#) definieron las funciones de pérdida aditivas CB-decomponibles de acuerdo a un CB-MBC tal que se cumple:

$$\lambda(\mathbf{c}', \mathbf{c}) = \sum_{j=1}^r \lambda_j \left( \mathbf{c}'^{\downarrow V_{C_{Comp_j}}}, \mathbf{c}^{\downarrow V_{C_{Comp_j}}} \right),$$

donde  $\lambda_j$  es una función de pérdida no negativa definida en  $I_j$ . Por ejemplo, los autores utilizan la distancia de Hamming en un ejemplo de funcionamiento de estas funciones. De esta manera, el problema de minimización para calcular la configuración de clases que minimiza la pérdida esperada se puede transformar en  $r$  sub-problemas de minimización sobre los distintos componentes conexos maximales del CB-MBC:

$$\arg \min_{\mathbf{c}' \in I} R(\mathbf{c}' | \mathbf{x}) = \left( \mathbf{c}^{*\downarrow V_{C_{Comp_1}}}, \dots, \mathbf{c}^{*\downarrow V_{C_{Comp_r}}} \right), \text{ donde}$$

$$\mathbf{c}^{*\downarrow V_{C_{Comp_j}}} = \arg \min_{\mathbf{c}'^{\downarrow V_{C_{Comp_j}}} \in I_j} \sum_{\mathbf{c}^{\downarrow V_{C_{Comp_j}}} \in I_j} \lambda_j \left( \mathbf{c}'^{\downarrow V_{C_{Comp_j}}}, \mathbf{c}^{\downarrow V_{C_{Comp_j}}} \right) \cdot \varnothing_j^{\mathbf{x}} \left( \mathbf{c}^{\downarrow V_{C_{Comp_j}}} \right).$$

## 2.4. Aprendizaje a partir de un conjunto de datos

En la literatura se han propuesto varios métodos para aprender la estructura del MBC que mejor representa un conjunto de datos (Tabla 2.3), mientras que ninguno de ellos aborda el problema de estimar los parámetros del modelo dado que este proceso se realiza como en las redes Bayesianas estándar. Se recuerda el crecimiento súper exponencial del espacio de estructuras de los MBCs, y por ello la necesidad de estos algoritmos de aprendizaje para que se muevan eficientemente por el espacio de estructuras. Existen dos aproximaciones cuando se quiere aprender una estructura de grafo desde un conjunto de datos: *score-based* o *score+search*, la cual trata de encontrar la estructura que maximice una puntuación, por ejemplo la verosimilitud penalizada de los datos dada la propia estructura, y *constrained-based*, la cual intenta encontrar una estructura que represente todas las independencias condicionales entre las variables.

### 2.4.1. Algoritmos *score-based*

La primera aproximación para aprender MBCs desde un conjunto de datos se propuso en [van der Gaag and de Waal \[2006\]](#). Los autores se centraron en aprender de manera eficiente estructuras de la familia árbol-árbol, tal que utilizando una técnica de aprendizaje *score+search* basada en el *score Minimum Description Length* [[Rissanen, 1978](#)], demuestran que los subgrafos clase y predictor se pueden aprender de manera óptima e independiente dado un subgrafo puente, ambos en tiempo polinómico. Primero, se aprende el subgrafo clase buscando el árbol ponderado no dirigido de recubrimiento máximo [[Kruskal, 1956](#)], donde el peso de cada arco es la información mutua entre el par de variables clase correspondiente, y el resultado se transforma en un

árbol dirigido utilizando el algoritmo de [Chow and Liu \[1968\]](#). Segundo, se aprende el subgrafo predictor para un subgrafo puente fijo buscando el árbol ponderado dirigido de recubrimiento máximo [[Chu and Liu, 1965](#)], donde el peso de cada arco es la información mutua condicionada entre el par de variables predictoras correspondiente dados los padres clase de la segunda predictora. Aunque estas dos técnicas de aprendizaje son *filter*, el subgrafo puente se va modificando vorazmente de manera *wrapper*, intentando mejorar el *global accuracy* (ec. 2.3), y por lo tanto resultando en un algoritmo híbrido.

Los mismos autores extienden este estudio en [de Waal and van der Gaag \[2007\]](#) a la familia de MBCs poliárbol-poliárbol. Los autores teorizan las condiciones requeridas para recuperar las estructuras poliárbol óptimas de los subgrafos clase y predictor. Si bien, este estudio teórico no expone cómo calcular el subgrafo puente.

En [Rodríguez and Lozano \[2008\]](#) se presenta un algoritmo novedoso para aprender MBCs con estructuras *kDB* en ambos subgrafos predictor y clase (un caso especial de MBC DAG-DAG). Los autores utilizan un algoritmo evolutivo, donde un individuo corresponde a una estructura MBC que se codifica como una ristra binaria en relación a la presencia o ausencia de cada arco posible. En concreto, utilizan una estrategia multi-objetivo mediante el algoritmo NSGA-II [[Deb et al., 2000](#)] tal que las funciones objetivo son el *accuracy*  $Acc_j$  sobre cada variable clase  $C_j$  (ec. 2.4). La salida del algoritmo es el conjunto eficiente de Pareto de las estructuras no dominadas, por lo que es necesario elegir aquella que resulte más conveniente para el problema en concreto.

[Bielza et al. \[2011\]](#) propusieron tres métodos para aprender MBCs DAG-DAG generales:

- Un algoritmo puramente *filter*, el cual se basa en la idea de [van der Gaag and de Waal \[2006\]](#) de resolver el problema de aprendizaje como dos problemas separados. Primero, se busca la mejor estructura para el subgrafo clase independientemente del resto de variables, lo cual se ejecuta una sola vez. Después, se busca la mejor estructura del subgrafo predictor condicionado a los padres clase dado un subgrafo puente fijo, el cual se va modificando para obtener aquel que consiga la mejor puntuación. A diferencia de en [van der Gaag and de Waal \[2006\]](#), el subgrafo puente se aprende de manera *filter* con una puntuación descomponible, tal que entre iteraciones sólo se modifica un arco del subgrafo y por lo tanto sólo se necesitan cálculos locales sobre las variables involucradas. Para el aprendizaje de los subgrafos predictor y clase se impone un orden ancestral entre las variables con el fin reducir el tiempo de cómputo. Los autores ponen de ejemplo el algoritmo K2 [[Cooper and Herskovits, 1992](#)] para el aprendizaje de estos subgrafos.

- Un algoritmo híbrido igual que el anterior puramente *filter*, pero el aprendizaje del subgrafo puente se guía por el *global accuracy*, o cualquier otra medida de rendimiento, en vez de por una puntuación *filter*.

- Un algoritmo *wrapper* que de manera voraz busca añadir o eliminar un arco en cualquier posición, pero respetando la estructura de un MBC, que mejore el *global accuracy* (ec. 2.3). El algoritmo termina cuando no se puede añadir o eliminar ningún arco de la estructura actual tal que se obtenga una mejora.

[Zaragoza et al. \[2011\]](#) también presentaron un algoritmo híbrido que consiste en una primera fase *filter* para aprender muy rápido una estructura inicial con las dependencias más fuertes, y después refinirla en una segunda fase *wrapper* más intensiva computacionalmente. En la primera fase se aprenden de manera independiente dos árboles como subgrafos clase y predictor mediante el recubrimiento ponderado máximo [[Chow and Liu, 1968](#)] con la información mutua como peso del arco entre dos variables. Después, se añaden aquellos arcos del subgrafo puente cuya información mutua entre la variable clase y predictora conectadas supere un umbral establecido. Finalmente, en la segunda fase se añaden aquellos arcos del subgrafo puente que consigan una mejora en el *accuracy* del MBC. Los autores experimentaron que establecer un umbral en la primera fase tal que se acepten sobre el 30% de los arcos en el subgrafo puente es una configuración que obtiene buenos resultados.

En el trabajo de [Antonucci et al. \[2013\]](#) se propone un modelo basado en *ensemble*, i.e., varios MBCs forman parte de un mismo modelo, tal que sus predicciones se conjuntan en una final. El subgrafo clase de cada MBC en el *ensemble* es un árbol que tiene como nodo raíz una variable clase  $C_j$  diferente, la cual se conecta a todas las demás sin que haya ninguna conexión entre éstas (siguiendo la condición de Markov, esto corresponde a asumir que el resto de clases son independientes dada  $C_j$ ). Las variables predictoras se asumen independientes dadas las clases, por lo que todos los clasificadores del *ensemble* son MBCs árbol-vacio. El subgrafo puente es el mismo para todos los MBCs, el cual se calcula con una aproximación *filter* maximizando la puntuación BDEu [[Buntine, 1991](#)] de cada variable predictora de manera independiente.

#### 2.4.2. Algoritmos *constrained-based*

[Qazi et al. \[2007\]](#) proponen un algoritmo para aprender MBCs DAG-vacio, aunque los autores no son conscientes de la definición formal en la literatura de los MBCs. En primer lugar, aprenden la estructura DAG del subgrafo clase utilizando procedimientos estándar de redes Bayesianas. Luego, obtienen el conjunto de variables predictoras más relevantes para cada variable clase utilizando el test de Kolmogorov-Smirnov [[Chakravarty et al., 1967](#), [Biesiada and Duch, 2005](#)]. Las predictoras seleccionadas se conectan a cada variable clase (subgrafo puente) asumiendo independencia condicionada con la clase, i.e., como en un modelo *naive Bayes*.

[Borchani et al. \[2012\]](#) se motivan por el hecho de que la clasificación de una variable

clase no se ve afectada por variables que no forman parte de su Markov blanket (MB), i.e., sus padres, hijos y esposos en la estructura  $G$ . Por ello, extienden el algoritmo HITON [Aliferis et al., 2010a,b] al contexto multi-dimensional para determinar el MB de cada variable clase, y después deducir de manera sencilla los subgrafos del MBC. Esta aproximación es escalable con respecto a la dimensión del conjunto de datos, ya que el MB de cada variable clase se puede aprender por separado. El algoritmo propuesto se extiende en Borchani et al. [2016] para hacer frente al cambio de concepto de los flujos de datos multi-dimensionales (Sección 5.2.1).

Ortigosa-Hernández et al. [2012] siguen una aproximación similar al incluir aquellos arcos entre dos variables que no superan un test de independencia, siempre y cuando la inclusión de estos arcos respete la estructura del MBC. Se sabe que  $2N \cdot MI(Z_i, Z_j)$  sigue asintóticamente una distribución  $\chi^2$  con  $(|\Omega_{Z_i}| - 1)(|\Omega_{Z_j}| - 1)$  grados de libertad si  $Z_i$  y  $Z_j$  son independientes, donde  $MI(C_i, C_j)$  es la información mutua entre dos variables cualesquiera  $Z_i$  y  $Z_j$  [Kullback, 1997]. Esto lo utilizan los autores para evaluar la independencia entre dos variables clase, y una variable clase y otra predictora, completando de esta manera los subgrafos clase y puente, respectivamente. También se sabe que  $2N \cdot MI(Z_i, Z_j | \mathbf{Z})$  sigue asintóticamente una distribución  $\chi^2$  con  $(|\Omega_{Z_i}| - 1)(|\Omega_{Z_j}| - 1) |\mathbf{Z}|$  grados de libertad si  $Z_i$  y  $Z_j$  son independientes dado un conjunto de variables  $\mathbf{Z}$ , donde  $MI(Z_i, Z_j | \mathbf{Z})$  es la información mutua entre  $Z_i$  y  $Z_j$  condicionado a  $\mathbf{Z}$  [Kullback, 1997]. Esta idea se aplica entre dos variables predictoras condicionadas a los padres clase de la segunda predictora para encontrar los arcos del subgrafo predictor. Por otro lado, los autores extienden el aprendizaje de los MBCs al escenario semi-supervisado (i.e., con muchos datos sin etiquetar) mediante una adaptación del algoritmo EM [Dempster et al., 1977] con bastante similaridad al algoritmo EM estructural Bayesiano de Friedman [1998].

Zhu et al. [2016] se motivan por el hecho de que un test de independencia sólo afirma si las variables son independientes o no, más que cuantificar el grado de su dependencia. Por ello, definen los coeficientes de dependencia entre dos variables  $Z_i$  y  $Z_j$  como  $c_{ij\alpha} = \min_{k \neq i,j} \{2N \cdot MI(Z_i, Z_j | Z_k) - \chi_{\alpha,l}\}$ , donde  $\chi_{\alpha,l}$  es el valor crítico para el nivel de significancia  $\alpha$  de una distribución  $\chi^2$  con  $l = (|\Omega_{Z_i}| - 1)(|\Omega_{Z_j}| - 1) |\Omega_{Z_k}|$  grados de libertad. Si  $c_{ij\alpha} > 0$ , existe una dependencia estadísticamente significativa entre las dos variables, y en caso contrario, hay al menos una manera de condicionar las variables tal que la dependencia no esté presente. De esta manera, los autores proponen un modelo híbrido basándose en estos coeficientes de dependencia, a la vez de utilizar una estrategia *score+search* que maximice la puntuación  $\sum_{i=1}^n \sum_{j=1, j \neq i}^n c_{ij\alpha} a_{ij}$  en el conjunto factible de estructuras que mantenga la topología restringida de un MBC, donde  $a_{ij} = 1$  si existe un arco entre las variables  $Z_i$  y  $Z_j$ , y 0 en caso contrario.

### 2.4.3. MBCs tratables

A diferencia de los anteriores métodos, sólo unos pocos trabajos en la literatura se han propuesto que consideren la complejidad de inferencia de los MBCs durante su proceso de aprendizaje. Por ejemplo, [Corani et al. \[2014\]](#) extienden su anterior trabajo [[Antonucci et al., 2013](#)] para aprender un único MBC bosque-vacío, también mediante una estrategia *filter* guiada por la puntuación BDEu. Aún con esta escasez de arcos en el modelo aprendido, la anchura de árbol de la estructura puede ser bastante grande y solamente acotada por el número de variables clase (propiedad ec. 2.12).

[Borchani et al. \[2010\]](#) introdujeron el primer método para aprender CB-MBCs, el cual se basa en una estrategia *wrapper*. Primero, se aprende un *naive Bayes* selectivo [[Langley and Sage, 1994](#)] para cada variable clase, y después se eliminan todas las predictoras hijas en común entre cualquier par de variables clase en base a dos criterios de ranking y *accuracy*, de tal manera que se obtiene un primer subgrafo puente con tantos componentes conexos maximales como variables clase. Segundo, se aprende el subgrafo predictor mediante la adición iterativa de aquellos arcos que consigan una mejora en *accuracy*. Esta fase se puede aprovechar del aspecto descomponible del MBC, dado que la adición de un arco hacia la variable  $X_j$  solamente cambiará el *accuracy* local del componente conexo maximal al que pertenece  $X_j$ . Finalmente, en una tercera fase se fusionan iterativamente los componentes en base a la adición del arco entre variables clase de diferentes componentes que consiga la mejor ganancia en *accuracy*, hasta que no haya ningún arco cuya inclusión haga que mejore dicho *accuracy* o no haya más componentes que fusionar, y consecuentemente en ese momento se actualizan los subgrafos puente y predictor del componente fusionado mediante la adición iterativa de arcos que también mejoren el *accuracy*. [Fernandez-Gonzalez et al. \[2015\]](#) adaptaron este mismo algoritmo para permitir trabajar con variables predictoras continuas que siguen una distribución Gaussiana sin la necesidad de discretizarlas, siendo éste el único trabajo que permite trabajar con este tipo de variables continuas en MBCs.

Sin embargo, ninguno de los trabajos anteriores proporciona garantías en la tratabilidad de la clasificación de los modelos aprendidos. [Pastink and van der Gaag \[2015\]](#) proponen un método que sí permite calcular la MPE en tiempo polinómico porque busca un MBC bosque-vacío que no supera una anchura de árbol fijada. Primero, aprenden un bosque como subgrafo clase que ya no se modifica hasta el final del algoritmo. Segundo, aprenden el subgrafo puente de manera *filter* con la puntuación BDEu y con una estrategia de búsqueda y poda. Para no sobrepasar la anchura de árbol fijada, ésta se calcula por cada estructura nueva candidata y se rechaza en caso de exceder la cota. La anchura se calcula sobre el grafo podado para reducir tiempo de cómputo (propiedad ec. 2.13). Opcionalmente en una tercera fase se permite obtener

una estructura de bosque como subgrafo predictor mediante la adición de arcos que mejoren la puntuación BDEu y no excedan la anchura de árbol.

Finalmente, Benjumeda et al. [2018] aprenden MBCs DAG-DAG generales mediante una adaptación de la búsqueda basada en orden [Bouckaert, 1992] tal que sólo se consideran aquellos órdenes en los que las variables clase preceden a las variables predictoras, y por lo tanto no se permiten arcos de predictoras a clases. Se aplica una estrategia voraz con cambios locales en los órdenes evaluados [Teyssier and Koller, 2005] junto con una lista tabú para reducir el tiempo de cómputo y reinicios aleatorios para evitar óptimos locales. Los autores proponen dos estrategias para garantizar la tratabilidad de los MBCs aprendidos. La primera, más costosa computacionalmente pero con mayor precisión predictiva, se basa en rechazar las estructuras cuya anchura de árbol del grafo podado supere una cota fijada. Se diferencia del anterior método de Pastink and van der Gaag [2015] porque no requiere que se tenga un subgrafo predictor vacío (explicado en la Sección 2.3.2). La segunda estrategia se basa en aprender CB-MBCs que no tengan ningún componente conexo maximal con más variables clase de una cota fijada. Esta estrategia es más eficiente dado que el cálculo del número de variables clase por componente es insignificante.

Tabla 2.3: Recopilación de los métodos propuestos en la literatura para aprender MBCs desde un conjunto de datos.

Referencia	Estrategia	Familia de MBC	Tratable
[van der Gaag and de Waal, 2006]	<i>Filter score-based</i>	árbol-árbol	No
[de Waal and van der Gaag, 2007]	<i>Filter score-based</i>	poliárbol-poliárbol	No
[Qazi et al., 2007]	<i>Constrained-based</i>	DAG-vacío	No
[Rodríguez and Lozano, 2008]	<i>Wrapper score-based</i>	DAG-DAG especial	No
[Borchani et al., 2010]	<i>Wrapper score-based</i>	CB-MBC	No
[Bielza et al., 2011]	<i>Filter</i> <i>Wrapper</i> Híbrido	<i>score-based</i> DAG-DAG	No
[Zaragoza et al., 2011]	Híbrido <i>score-based</i>	árbol-árbol	No
[Borchani et al., 2012]	<i>Constrained-based</i>	DAG-DAG	No
[Ortigosa-Hernández et al., 2012]	<i>Constrained-based</i>	DAG-DAG	No
[Antonucci et al., 2013]	<i>Filter score-based</i>	árbol-vacío	No
[Corani et al., 2014]	<i>Filter score-based</i>	bosque-vacío	No
[Fernandez-Gonzalez et al., 2015]	<i>Wrapper score-based</i>	CB-MBC	No
[Pastink and van der Gaag, 2015]	<i>Filter score-based</i>	bosque-{vacío,bosque}	Sí
[Zhu et al., 2016]	<i>Filter</i> híbrido	DAG-DAG especial	No
[Benjumeda et al., 2018]	<i>Filter score-based</i>	DAG-DAG	Sí

## 2.5. Aplicaciones en la literatura

La primera aplicación abordada con un MBC fue un problema médico por [Qazi et al. \[2007\]](#). Si bien, los autores no eran conscientes de la existencia de una definición formal en la literatura del modelo que estaban utilizando [[van der Gaag and de Waal, 2006](#)]. Después, no han surgido muchos problemas que se hayan modelado con un MBC. La mayoría de ellos también se relacionan con problemas médicos, mientras que sólo unos pocos se encuadran en otros dominios. A continuación se describe un resumen de las aplicaciones encontradas en la literatura que se han abordado con MBCs. En la Tabla 2.4 se compila un listado de las mismas.

### 2.5.1. Problemas médicos

Muchos dominios médicos incluyen problemas de clasificación multi-dimensional: un gen puede tener múltiples funciones biológicas, un paciente puede sufrir de múltiples enfermedades, un paciente puede volverse resistente a múltiples medicamentos para un tratamiento, etc. Se pueden destacar las siguientes aplicaciones:

- *Diagnóstico de la enfermedad de las arterias coronarias* mediante la predicción de anomalías en el movimiento de las paredes de los 16 segmentos en los que se divide el ventrículo izquierdo del corazón [[Qazi et al., 2007](#)]. Cada clase binaria, i.e., cada uno de los segmentos, se predice como normal o anormal, por lo que el problema se encuadra en una configuración multi-etiqueta. El conjunto de datos con el que se trabaja está etiquetado con hasta cuatro diferentes tipos de anomalías, por lo que se podría evitar la simplificación que han hecho los autores de juntar indistintivamente en una sola clase todos los tipos de anomalías, y tratar el problema como clasificación multi-dimensional.
- *Estimación de la calidad de vida* relacionada con la salud de pacientes que padecen la enfermedad de Parkinson [[Borchani et al., 2012](#)]. Las cinco variables clase (movilidad, auto-suficiencia, actividades habituales, dolor/malestar y ansiedad/depresión) tienen tres valores de respuesta: sin problemas, algunos problemas y problemas severos.
- *Asistencia en el tratamiento de la esclerosis múltiple* mediante la predicción del subtipo de enfermedad de cuatro posibles y el tiempo esperado para alcanzar un nivel de gravedad que indique que se requiere asistencia para caminar [[Rodríguez et al., 2012](#)]. Para enfrentarse a este problema en un contexto de clasificación supervisada, esta última variable clase se ha discretizado en cuatro intervalos.
- *Predicción de los inhibidores del virus de la inmunodeficiencia humana tipo 1* (VIH-1), tanto con inhibidores de la transcriptasa inversa (ITIs) como de la proteasa (IPs) [[Borchani et al., 2013](#)]. Se consideran diez y ocho medicamentos,

respectivamente, de cada uno de los tipos de inhibidores anteriores. Se trata de un problema multi-etiqueta ya que un paciente es resistente o no a un medicamento.

- *Clasificación de neuronas* [Fernandez-Gonzalez et al., 2015]. El objetivo es determinar su especie (rata, humano, ratón o elefante), sexo (macho o hembra), tipo de célula nivel uno (neurona principal o interneurona), tipo de célula nivel dos de seis posibles valores, fase de desarrollo (recién nacido, joven, adulto o anciano) y la región del cerebro de catorce posibles donde se encuentra.

### 2.5.2. Otras aplicaciones

Finalmente, se enumeran otras aplicaciones donde se han aplicado MBCs:

- *Análisis de sentimientos* mediante la caracterización de la actitud de un cliente en su valoración basándose en tres variables objetivo: subjetividad (objetiva o subjetiva), polaridad del sentimiento (muy negativa, negativa, neutral, positiva y muy positiva) y voluntad de influir (texto declarativo, fuerte voluntad, mediana y suave) [Ortigosa-Hernández et al., 2012].
- *Predicción del reclutamiento de peces* [Fernandes et al., 2013]. Se estudian tres especies de interés comercial en el golfo de Vizcaya: anchoa, sardina y merluza. Para cada especie, los autores dividían el reclutamiento en bajo, medio y alto.

Tabla 2.4: Problemas de clasificación multi-dimensional encontrados en la literatura que se han modelado con un MBC. El número de variables predictoras,  $m$ , corresponde tras haber aplicado un proceso de selección de variables. Siguiendo la terminología expuesta,  $d$  es el número de variables clase e  $|I|$  el espacio de configuraciones.

Referencia	Problema de clasificación	Dimensión			
		$m$	$d$	$ I $	
[Qazi et al., 2007]	Enfermedad coronaria	108	16	65536	
[Borchani et al., 2012]	Calidad de vida	14	5	243	
[Ortigosa-Hernández et al., 2012]	Análisis de sentimientos	14	3	40	
[Rodríguez et al., 2012]	Esclerosis múltiple	21	2	16	
[Borchani et al., 2013]	Inhibidores del VIH-1	$\begin{cases} \text{ITIs} \\ \text{IPs} \end{cases}$	25	10	1024
			55	8	256
[Fernandes et al., 2013]	Reclutamiento de peces	15-138	3	27	
[Fernandez-Gonzalez et al., 2015]	Neuroanatomía	81	6	5376	

# Capítulo 3

## Propuesta de un nuevo clasificador multi-dimensional en árbol

### 3.1. Definición y contexto

Los meta-clasificadores combinan diferentes modelos antes de realizar una predicción motivados por el *no free-lunch theorem* [Wolpert and Macready, 1997], el cual establece que no existe un único algoritmo de aprendizaje que siempre induzca el clasificador más preciso en cualquier dominio. Un clasificador híbrido es un meta-clasificador que se induce teniendo en cuenta dos o más paradigmas. Un ejemplo de clasificador híbrido es el *naive Bayes tree* (NBTree) de Kohavi [1996], el cual sitúa un modelo NB en cada nodo hoja de un árbol de clasificación. Landwehr et al. [2005] sigue una idea similar con el *logistic model tree*, el cuál coloca regresiones logísticas en vez de NBs en las hojas del árbol de clasificación. Un ejemplo más de clasificador híbrido es el *lazy Bayesian rules* propuesto por Zheng and Webb [2000], el cual construye una regla con un clasificador NB en su consecuente por cada ejemplo de prueba.

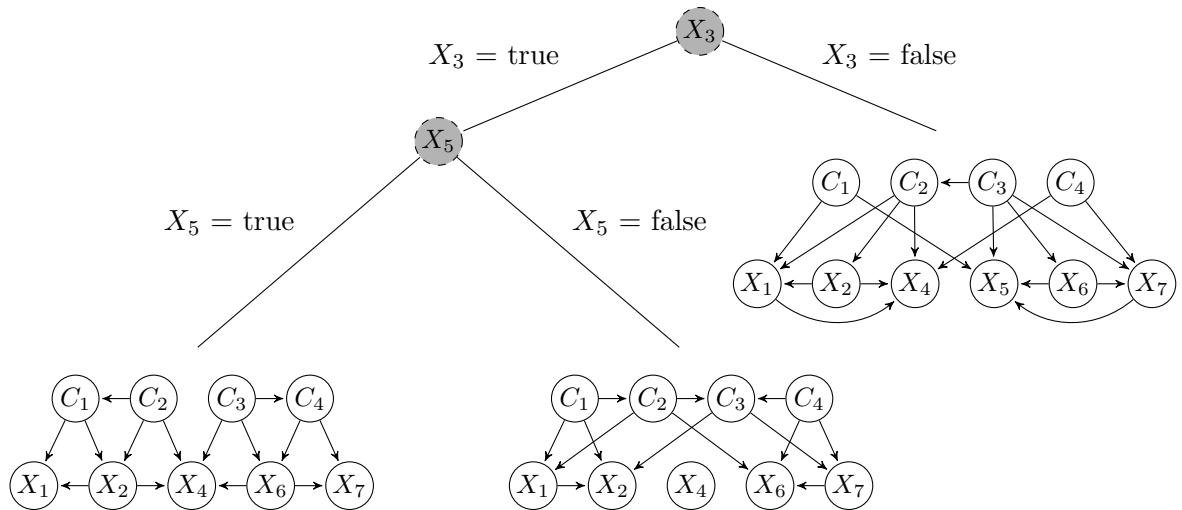


Figura 3.1: Ejemplo de la estructura de un MBCTree.

En este trabajo proponemos un híbrido de árboles de clasificación [Breiman et al., 1984] y MBCs. Hasta el mejor de nuestro conocimiento, este modelo es el primer híbrido propuesto en el contexto de clasificación multi-dimensional, el cual denominamos *multi-dimensional Bayesian network classifier tree* (MBCTree). Un MBCTree es un árbol de clasificación con MBCs en las hojas (Figura 3.1):

- Un nodo interno de un MBCTree corresponde a una variable predictora  $X_i$  como en los árboles de clasificación estándar, y se bifurca en tantas ramas como posibles valores de  $X_i$ , i.e.,  $|\Omega_{X_i}|$ . Las ramas están etiquetadas con los posibles valores de  $X_i$ .
- Un nodo hoja de un MBCTree es un MBC sobre todas las variables clase y sobre aquellas variables predictoras que no están presentes en el camino desde la raíz del árbol hasta la propia hoja. Un MBCTree puede ser asimétrico, por lo que los MBCs de las hojas pueden tener diferentes variables predictoras. Una nueva instancia se clasificará haciéndola descender por el árbol desde la raíz hasta un MBC hoja de acuerdo a los resultados de las pruebas a lo largo del camino.

### 3.2. Algoritmo de aprendizaje *wrapper*

Proponemos una aproximación *wrapper* guiada por el *global accuracy* (ec. (2.3)) para aprender MBCTrees desde un conjunto de datos, aunque cualquier otra medida de evaluación de rendimiento definida en la Sección 2.2 se podría utilizar de manera indiferente. La aproximación que proponemos se detalla en el Algoritmo 1.

El MBCTree se aprende mediante la elección recursiva como nodo interno de la variable predictora  $X_i$  que realiza la mejor partición de los datos, i.e., que consigue el *global accuracy* más alto [pasos 1-10], hasta que separar los datos ya no agrega valor a las predicciones [pasos 11-13]. La aproximación propuesta se encuadra como un algoritmo voraz de arriba hacia abajo [Quinlan, 1986]. Se empieza en el nodo raíz del árbol y se calcula el *global accuracy*,  $Acc_{MBCTree_i}$ , de la partición por cada variable predictora  $X_i$  [pasos 2-8]. Para ello, se entrena un MBC por cada valor posible de  $X_i$  con su porción de datos correspondiente [pasos 3-6]. El *global accuracy* de la partición se calcula descendiendo el conjunto de datos de prueba a los MBCs aprendidos en base a los valores de  $X_i$  [paso 7]. Hemos elegido una estrategia *wrapper* guiada por el *global accuracy* para aprender los MBCs como en Bielza et al. [2011]. Este proceso se repite sobre cada subconjunto derivado de la mejor partición de manera recursiva [paso 10], hasta que no haya ninguna partición que mejore el *global accuracy* conseguido por un MBC aprendido con los datos que alcanzan el nodo actual [paso 1]. En tal caso, este MBC se sitúa como un nodo hoja [paso 12]. También se debe crear un nodo hoja si solamente queda una variable predictora para separar los datos, dado que no podría

existir un MBC en la hoja sin variables predictoras en su estructura, y también si no hay suficientes datos para seguir creciendo el árbol, i.e., para aprender los MBCs de las particiones. Para evitar sobre-ajustarse al conjunto de entrenamiento, se puede exigir una mínima mejora en el paso 9 como estrategia de poda, tal que la recurrencia termina en el paso 12 si no se consigue ninguna mejora significativa.

---

**Algoritmo 1** Algoritmo *wrapper* para aprender MBCTrees desde un conjunto de datos.

---

Entrada: Un conjunto de datos etiquetado  $D$

Salida: Un MBCTree entrenado con el conjunto de datos  $D$

- 1: Aprender un MBC y calcular su *global accuracy*,  $Acc_{MBC}$ , con los datos  $D$
  - 2: **para cada** variable predictora  $X_i$  **hacer**
  - 3:     Separar  $D$  en  $|\Omega_{X_i}|$  subconjuntos  $D_{ij}$  en base a los posibles valores  $j$  de  $X_i$
  - 4:     **para cada** suconjunto  $D_{ij}$  **hacer**
  - 5:         Aprender un  $MBC_{ij}$  con el subconjunto  $D_{ij}$
  - 6:     **fin para**
  - 7:     Calcular el *global accuracy*,  $Acc_{MBCTree_i}$ , de la partición en  $X_i$
  - 8:     **fin para**
  - 9:     **si**  $\exists Acc_{MBCTree_i} > Acc_{MBC}$  **entonces**
  - 10:         Crear un nodo interno  $X_i$  (con el  $Acc_{MBCTree_i}$  más alto). Para cada nodo hijo bajo una rama con etiqueta  $j \in \Omega_{X_i}$ , llamar recursivamente al algoritmo, desde el paso 1, sobre el subconjunto  $D_{ij}$
  - 11:     **si no**
  - 12:         Crear un nodo hoja con el MBC aprendido en el paso 1
  - 13:     **fin si**
- 

### 3.3. Estudio experimental

Con el fin de evaluar nuestro modelo y algoritmo de aprendizaje propuestos hemos realizado un estudio experimental sobre conjuntos de datos sintéticos. Este estudio sigue los pasos que se muestran en la Figura 3.2:

- En primer lugar, se genera un MBCTree aleatorio con una profundidad definida. Las variables predictoras que están asociadas a los nodos internos del árbol se eligen de manera aleatoria. Los MBCs hoja también se generan aleatoriamente, tal que los subgrafos clase y predictor son muestras uniformemente distribuidas de DAGs [Ide and Cozman, 2002] y cada variable clase  $C_j$  se conecta con cada variable predictora  $X_i$  con probabilidad  $p$ . Elegimos  $p = 0,5$  para tomar muestras de manera uniforme del espacio de estructuras de MBCs. Los parámetros de los MBCs se fuerzan a ser extremos, i.e., menores que 0.3 y mayores que 0.7.
- En segundo lugar, se simula un conjunto de datos del MBCTree generado. Para ello, se simula un subconjunto de datos de manera aleatoria por cada MBC hoja del MBCTree mediante el método *probabilistic logic sampling* [Henrion, 1988]. Se impone

que cada subconjunto contribuya en al menos un porcentaje definido al conjunto de datos total que se está simulando.

- En tercer lugar, se aprenden un MBC y un MBCTree siguiendo las aproximaciones *wrapper* en [Bielza et al. \[2011\]](#) y Algoritmo 1, respectivamente. Después, éstos se comparan en términos de precisión predictiva utilizando el conjunto de datos simulado, el cual se divide en un subconjunto de entrenamiento que contiene el 80 % de los ejemplos y un subconjunto de prueba con el 20 % restante. El conjunto de entrenamiento también se divide en los mismos porcentajes para aprender el MBCTree: el 80 % de las instancias se utiliza para aprender los MBCs y el 20 % para evaluarlos tal que se pueda calcular qué variable predictora hace la mejor partición. Se puede aplicar perfectamente una técnica de validación cruzada en ambos casos, pero hemos seguido esta estrategia de entrenar y evaluar en conjunto con una mayor cantidad de datos debido a aspectos de eficiencia computacional.

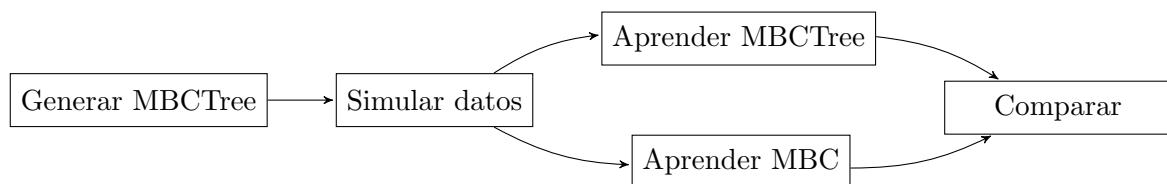


Figura 3.2: Procedimiento del estudio experimental realizado con datos sintéticos para evaluar el modelo MBCTree.

Hemos establecido dos configuraciones diferentes siguiendo el esquema comentado. Cada configuración se ha ejecutado diez veces, lo que ha conducido a los resultados que se muestran en la Tabla 3.1.

- Configuración A: se genera un MBCTree de profundidad 1 con 11 variables predictoras en los MBCs hoja (11 + 1 predictoras en conjunto) y 4 variables clase para predecir. Todas las variables son binarias. El conjunto de datos simulado consiste en 100.000 instancias, con al menos 20 % por cada rama.
- Configuración B: se genera un MBCTree de profundidad 2 con 10 variables predictoras en los MBCs hoja (10 + 2 predictoras en conjunto) y 4 variables clase para predecir. Todas las variables son binarias. El conjunto de datos simulado consiste en 100.000 instancias, con al menos 20 % por cada rama de manera recursiva.

### 3.3.1. Resultados obtenidos

Como resultado favorable del estudio experimental realizado, el modelo MBCTree que proponemos ha obtenido *global accuracies* más altos respecto a los MBCs en todas las ejecuciones de ambas configuraciones. El NBTree también obtenía mejoras similares respecto a los clasificadores NB estándar [[Kohavi, 1996](#)]. La mejora obtenida

se pronuncia ligeramente más en las ejecuciones de la Configuración B porque los datos simulados provienen de más distribuciones de probabilidad diferentes, y el MBCTree es capaz de descubrir estas particiones de los datos. En ambos casos, la mejora obtenida es estadísticamente significativa con un  $p$ -valor = 0,001953 cuando se utiliza la prueba de los rangos con signos de Wilcoxon.

Como aspecto positivo del enfoque de aprendizaje propuesto, el nodo raíz del MBCTree inicial que se genera de manera aleatoria siempre se incluye como nodo interno del MBCTree posteriormente aprendido. Si bien, esta inclusión no es siempre como nodo raíz, especialmente en las ejecuciones de la Configuración B. Este hecho se explica por la naturaleza voraz del algoritmo de aprendizaje propuesto.

Por el contrario, se debe destacar la mayor carga computacional que implica el aprendizaje de un MBCTree desde un conjunto de datos. El punto crítico es calcular para cada nodo interno qué variable predictora separa mejor los datos, ya que se tiene que entrenar un MBC por cada valor posible de cada predictora. En media, un total de 103 y 137 MBCs han sido necesarios entrenar para inducir los MBCTrees de las Configuraciones A y B, respectivamente. Nuestro método está añadiendo dos órdenes de magnitud para este problema particular. Tal complejidad se podría aliviar ligeramente si los MBCs se entrenaen de manera paralela. Otro inconveniente de nuestro modelo es que un MBCTree necesita un conjunto de datos de entrenamiento más grande porque las particiones recursivas de los mismos durante el proceso de aprendizaje causan que los MBCs hoja se entrenen con menos datos.

Tabla 3.1: Comparación en términos de *global accuracy* de los modelos MBC y MBCTree siguiendo el experimento sobre datos sintéticos generados aleatoriamente.

Configuración A				Configuración B			
MBC	MBCTree	Dif.	MBCs apr.	MBC	MBCTree	Dif.	MBCs apr.
0.7713	0.7783	+0.0070	79	0.7122	0.7240	+0.0118	79
0.7857	0.8032	+0.0175	59	0.7810	0.7986	+0.0176	59
0.8038	0.8145	+0.0107	215	0.7127	0.7292	+0.0165	139
0.7542	0.7716	+0.0174	121	0.7623	0.7793	+0.0170	159
0.7325	0.7407	+0.0082	79	0.7998	0.8106	+0.0108	79
0.7812	0.7905	+0.0093	121	0.5326	0.5503	+0.0177	59
0.7605	0.7771	+0.0166	59	0.7399	0.7553	+0.0154	274
0.7911	0.8102	+0.0191	157	0.7922	0.8036	+0.0114	79
0.7753	0.7859	+0.0106	59	0.7559	0.7753	+0.0194	249
0.7782	0.7878	+0.0096	79	0.7779	0.8089	+0.0310	197
<i>Media</i>		<b>+0.0126</b>	<b>103</b>	<i>Media</i>		<b>+0.0169</b>	<b>137</b>



# Capítulo 4

## Aplicaciones en la Industria 4.0

### 4.1. Definición del problema de clasificación

En este capítulo abordamos el problema de eficiencia energética sobre una máquina de mecanizado de una planta industrial de la empresa española Exte-Tar S.A. El objetivo consiste en determinar qué elementos activos de la máquina están encendidos en un momento dado. En una segunda aproximación al problema, el objetivo será determinar el grado de consumo en el que se encuentra cada uno. Esta información será de gran utilidad para en un futuro poder sincronizar de manera inteligente las puestas en marcha y paradas de los elementos activos con su consecuente ahorro energético.

La máquina de mecanizado sobre la que vamos a trabajar, expuesta en la Figura 4.1a, produce piezas de cigüeñales que posteriormente exporta a grandes empresas de la automoción como Ford o Jaguar. Para hacernos una idea de la dimensión de esta máquina, puede llegar a producir 1.000 unidades de cigüeñales por día. Aunque son varios los elementos activos que conforman el funcionamiento completo de la máquina, solamente tres son los grandes consumidores de energía siguiendo la regla del 80/20, i.e., sólo unos pocos elementos consumen el 80 % de la energía total. Estos tres elementos son el electromandrino o *spindle* que hace rotar a la herramienta para poder llevar a cabo el mecanizado, y los servos de los ejes X e Y que hacen mover la punta de la herramienta en cada eje. En un primer contacto con este problema se va a trabajar solamente sobre los grandes consumidores de la máquina, i.e., nuestro problema de clasificación consta de 3 variables clase, una por cada elemento activo comentado.

Los posibles valores de las variables clase varían en base a las dos aproximaciones que realizamos. La primera es una aproximación multi-etiqueta, tal que cada variable clase es binaria en relación a los dos posibles valores de encendido o apagado del elemento. En una segunda aproximación más general, el estado de encendido se divide en tres estados adicionales, tal que cada uno representa un grado de consumo del elemento. De esta manera, cada variable clase toma cuatro posibles valores ordinales: 0, el elemento

no está consumiendo; 1, el elemento está en mínimo rendimiento; 2, el elemento está a una potencia intermedia; y 3, el elemento se encuentra en máxima ejecución.



(a) Máquina de mecanizado.



(b) Sensor capturador de datos de consumo.

Figura 4.1: Máquina de mecanizado sobre la que enfocamos el problema de eficiencia energética. Fuente: Etxe-Tar S.A.

Los expertos de la empresa consiguen capturar datos del consumo total de la máquina mediante el sensor Oberon Energy (Figura 4.1b). Este sensor es capaz de trabajar a una frecuencia de 4kHz, i.e., durante un segundo es capaz de darnos cuatro mil instantes con sus respectivos datos de consumo. El conjunto de datos sobre el que vamos a trabajar consta de una hora de trabajo de la máquina, por lo que en este intervalo el sensor nos proporciona cerca de quince millones de instantes. De nuevo ante un primer contacto con este problema, vamos a simplificarlo y simular una menor frecuencia del sensor. Tras un estudio realizado, hemos observado que una reducción de la frecuencia de captura entre cien nos ofrece aparentemente la misma información, dado que las evoluciones de las variables de consumo a lo largo del tiempo resultan ser exactamente iguales para ambas frecuencias. Incluso se consigue mantener los picos de consumo, los cuales serían las características más propensas a desaparecer ante una merma en la frecuencia de captura. Se cuestiona de esta manera la necesidad de una frecuencia de captura tan elevada, y se propone como trabajo futuro la búsqueda de la frecuencia ideal para la captura de los datos de consumo.

Las variables predictoras de nuestro problema de clasificación con las que queremos predecir el estado de los elementos activos de la máquina son los niveles de consumo totales de ésta. En total, el sensor captura cinco variables de consumo, cada una en un sistema trifásico, por lo que resultan finalmente en quince variables predictoras:

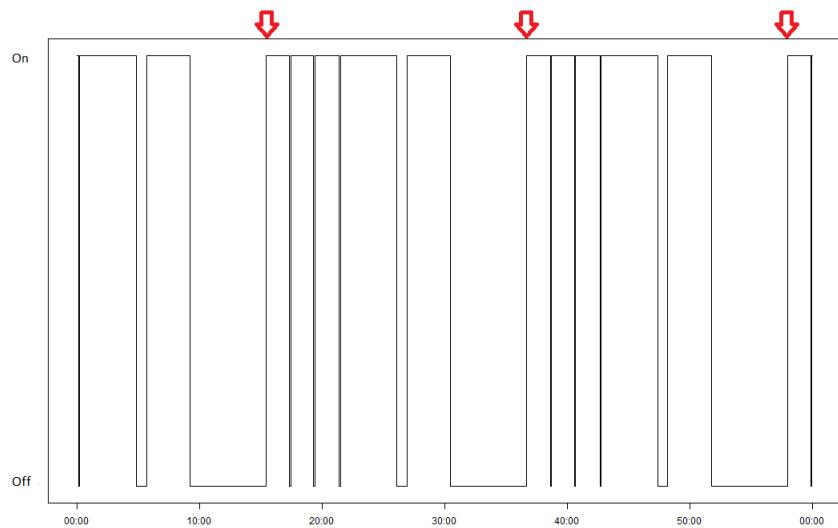
- Ia.A: Corriente fase a medida en amperios.
- Ib.A: Corriente fase b medida en amperios.
- Ic.A: Corriente fase c medida en amperios.
- Pa.W: Potencia activa fase a medida en vatios.

- Pb.W: Potencia activa fase b medida en vatios.
- Pc.W: Potencia activa fase c medida en vatios.
- Qa.var: Potencia reactiva fase a medida en voltiamperios reactivos.
- Qb.var: Potencia reactiva fase b medida en voltiamperios reactivos.
- Qc.var: Potencia reactiva fase c medida en voltiamperios reactivos.
- Sa.VA: Potencia aparente fase a medida en voltiamperios.
- Sb.VA: Potencia aparente fase b medida en voltiamperios.
- Sc.VA: Potencia aparente fase c medida en voltiamperios.
- Va.V: Voltaje fase a medida en voltios.
- Vb.V: Voltaje fase b medida en voltios.
- Vc.V: Voltaje fase c medida en voltios.

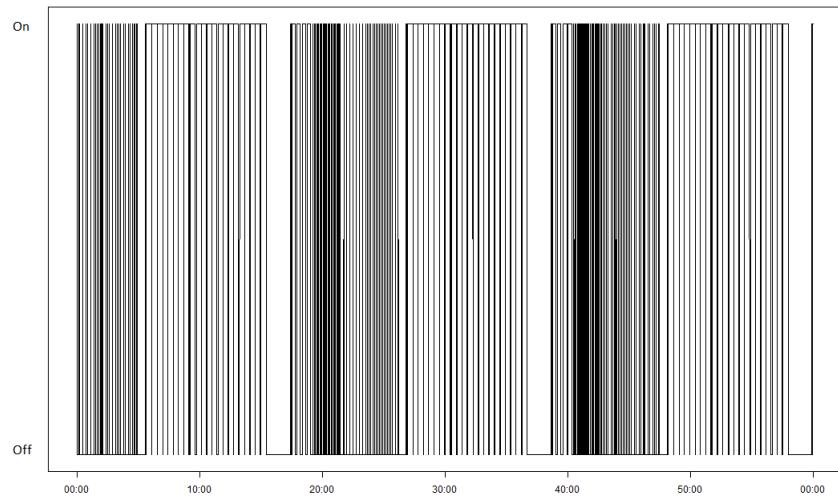
Las mediciones se han hecho durante ciclos en vacío de auto-diagnóstico de la máquina, i.e., no en ciclos de mecanizado. De esta manera, los expertos de la empresa son capaces de etiquetar las tres variables clase de cada instante capturado por el sensor en base a las revoluciones de los elementos activos. En conclusión, tenemos un conjunto de datos a los que a cada instante de consumo total de la máquina conocemos los valores verdaderos de cada elemento activo, por lo que nos encontramos ante un problema de clasificación supervisada multi-dimensional donde queremos aprender un modelo que asocie los niveles totales de consumo de la máquina con las distintas configuraciones de los estados de sus elementos activos, y de esta manera ser capaces de predecir los futuros estados de los elementos en base a los valores del consumo total de la máquina.

## 4.2. Aproximación multi-etiqueta

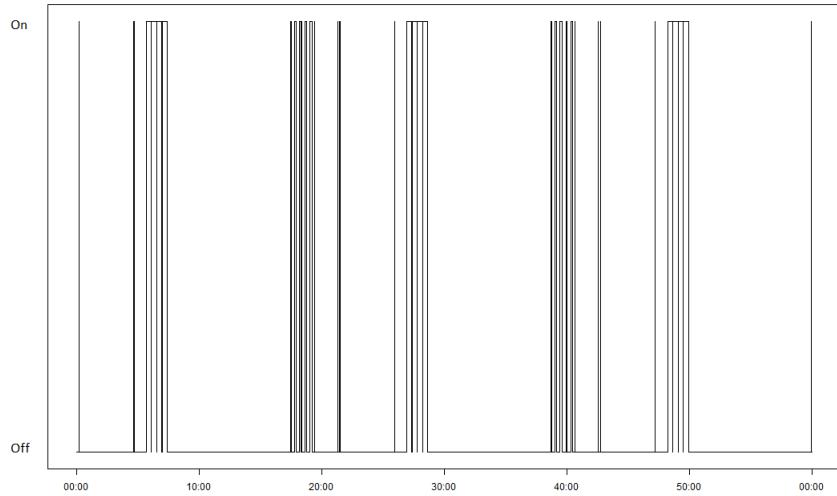
En una primera aproximación al problema queremos predecir únicamente si los elementos activos están encendidos o apagados en base al consumo total de la máquina. En la Figura 4.2 se muestra la evolución de las tres variables clase asociadas a los grandes consumidores de la máquina durante el intervalo de una hora del conjunto de datos sobre el que trabajamos. Cada elemento activo está encendido con un valor *On*, y apagado con un valor *Off*. En las gráficas se puede observar, por ejemplo, que el servo del eje X tiene mucho más trabajo que el servo del eje Y. Un dato importante es el porcentaje de tiempo que cada elemento está encendido, dado que establece una referencia para saber si nuestro modelo tiene un buen rendimiento. Es decir, un modelo ingenuo que siempre diga que el *spindle* está encendido va a conseguir un 63,74 % de *accuracy* sin ni siquiera aplicar ningún proceso de aprendizaje automático.



(a) Variable clase Spindle. 63,74 % encendido.



(b) Variable clase Eje X. 50,72 % encendido.



(c) Variable clase Eje Y. 10,58 % encendido.

Figura 4.2: Evolución temporal de los estados de encendido y apagado de los tres elementos activos del problema de clasificación de eficiencia energética.

Otro dato interesante de la anterior Figura 4.2 son los ciclos programados que realiza la máquina y que se reflejan en la secuencia de los estados de sus elementos activos. Se observa que a partir de las flechas rojas situadas en la parte superior de la Figura 4.2, los elementos activos repiten la misma sucesión de estados. En total, se contemplan dos ciclos completos durante la hora que dura el conjunto de datos, y otros dos ciclos sin empezar y sin acabar al principio y final del mismo, respectivamente. Nos aprovechamos de esta situación para entrenar nuestros modelos de clasificación con los datos hasta el final del primer ciclo completo ( $\sim$  minuto 37), y evaluarlos con todos los datos posteriores. De esta manera, estaremos entrenando los modelos con un ciclo completo y evaluando su rendimiento con los datos de otro ciclo completo futuro.

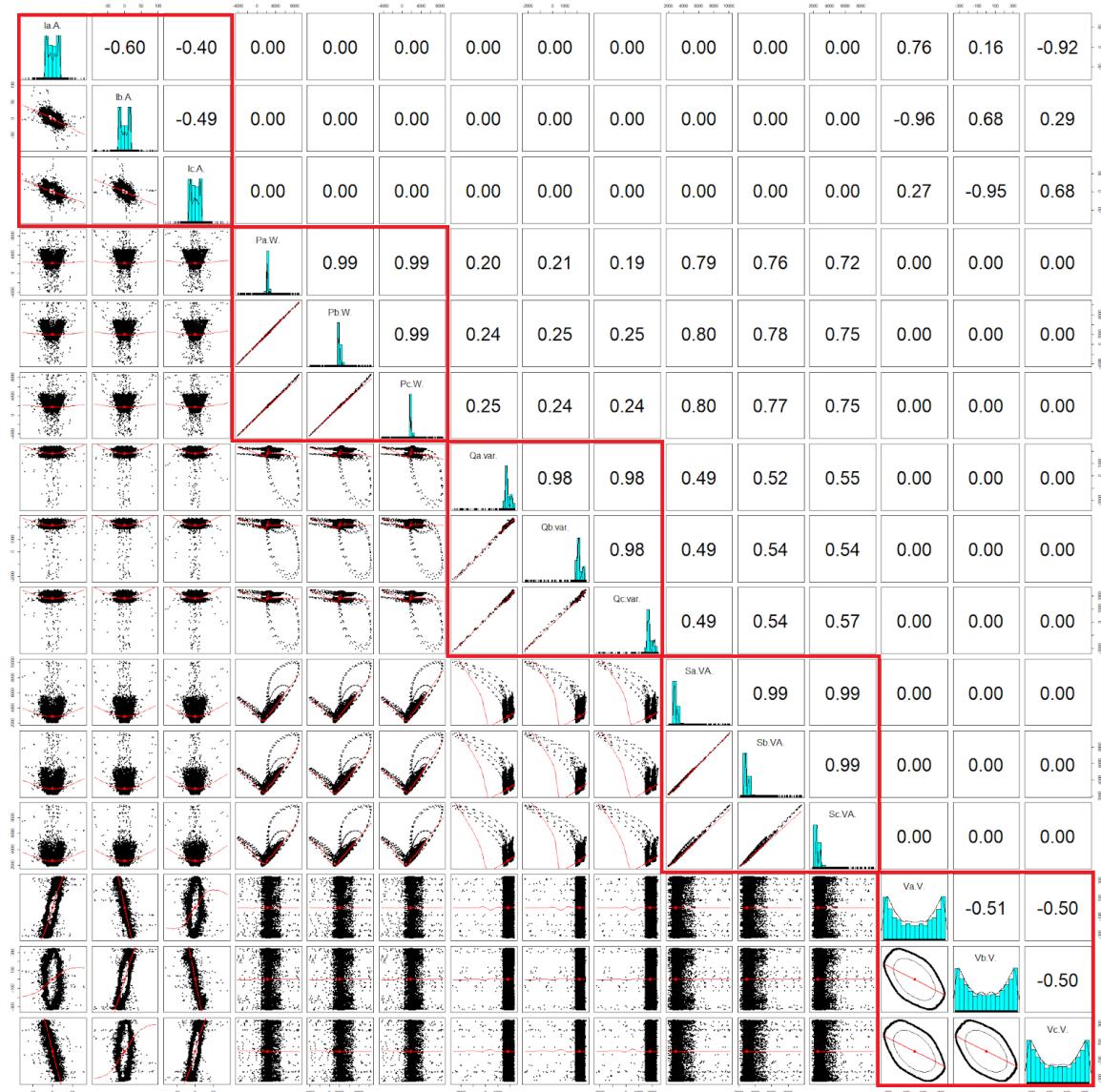


Figura 4.3: Detalle de las variables predictoras del problema de clasificación de eficiencia energética. En la diagonal, los histogramas de cada variable predictor. Por debajo, los diagramas de dispersión entre cada par de variables. Por encima, sus correlaciones de Pearson. En color rojo se agrupan las tres fases de una misma variable de consumo.

Antes de resolver el problema de clasificación es interesante realizar un estudio sobre las relaciones entre las variables del problema. En la Figura 4.3 se detalla un estudio sobre las variables predictoras, todas ellas de naturaleza continua, del que se extraen las correlaciones entre cada par de ellas. Se observa que las tres predictoras en relación a las tres fases de cada variable de consumo tienen una alta correlación entre ellas.

Otro estudio de gran interés es evaluar la relación de cada variable con respecto a las variables clase del problema, de tal manera que podamos concluir cuáles de ellas son de utilidad para la clasificación. Este proceso se conoce como selección de variables (en inglés, *feature subset selection*), y se basa en seleccionar aquellas variables predictoras más relevantes para entrenar el modelo de clasificación, ya que no existe una monotonía entre el número de variables utilizadas y el rendimiento que se consigue. Además, la selección de variables permite conseguir otras características deseables como la obtención de modelos más interpretables, la reducción de los tiempos de entrenamiento e inferencia, y una mayor generalización al evitar sobre-entrenar los datos de entrenamiento. Hemos abordado la selección de variables mediante la información mutua entre cada variable del problema y cada variable clase, obteniendo los resultados que se exponen en la Tabla 4.1.

Tabla 4.1: Información mutua entre cada una de las variables (filas) del problema de eficiencia energética y las tres variables clase (columnas) para la aproximación multi-etiqueta. En color rojo se remarcán aquellas parejas de variables con una información mutua baja.

	Spindle	Eje X	Eje Y
Ia.A	0.018	0.025	0.003
Ib.A	0.020	0.032	0.002
Ic.A	0.019	0.031	0.002
Pa.W	0.311	0.140	0.162
Pb.W	0.288	0.191	0.184
Pc.W	0.264	0.176	0.192
Qa.var	0.117	0.286	0.044
Qb.var	0.104	0.238	0.029
Qc.var	0.157	0.234	0.047
Sa.VA	0.112	0.206	0.143
Sb.VA	0.138	0.212	0.106
Sc.VA	0.134	0.226	0.109
Va.V	0.001	0.001	0.001
Vb.V	0.001	0.001	0.001
Vc.V	0.001	0.001	0.001
Spindle	-	0.148	0.071
Eje X	0.148	-	0.028
Eje Y	0.071	0.028	-

La información mutua entre dos variables  $X$  e  $Y$  mide la reducción de la incertidumbre de la variable  $X$  una vez se conoce el valor de  $Y$ :

$$MI(X, Y) = H(X) - H(X|Y),$$

donde  $H(X)$  es la entropía de Shannon de la variable aleatoria discreta  $X$ , i.e., la información esperada de sus eventos:

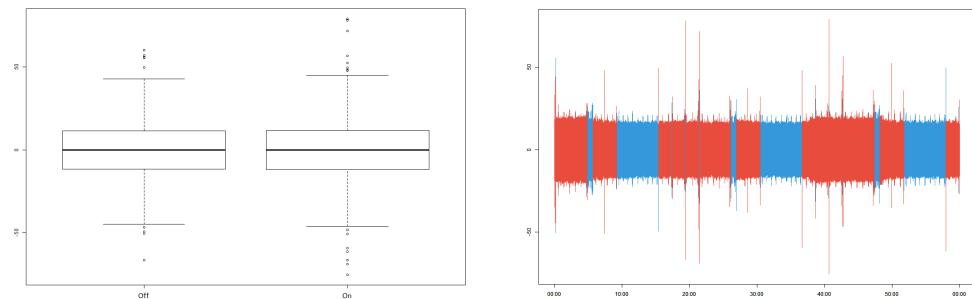
$$H(X) = - \sum_{i=1}^{|\Omega_X|} p(x_i) \log_2 p(x_i),$$

y  $H(X|Y)$  la entropía de la variable  $X$  condicionada a la variable  $Y$ :

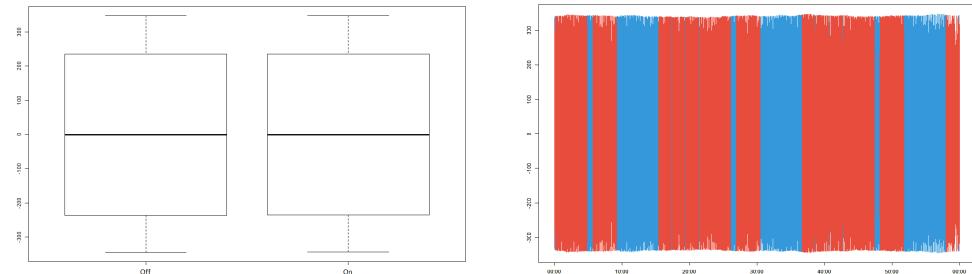
$$H(X|Y) = \sum_{j=1}^{|\Omega_Y|} p(y_j) H(X|Y = y_j) = - \sum_{i=1}^{|\Omega_X|} \sum_{j=1}^{|\Omega_Y|} p(x_i, y_j) \log_2 p(x_i|y_j).$$

Para calcular estas medidas se necesita que ambas variables sean discretas, por lo que las variables predictoras continuas se han discretizado mediante el método supervisado basado en el *minimum description length* de [Fayyad and Irani \[1993\]](#).

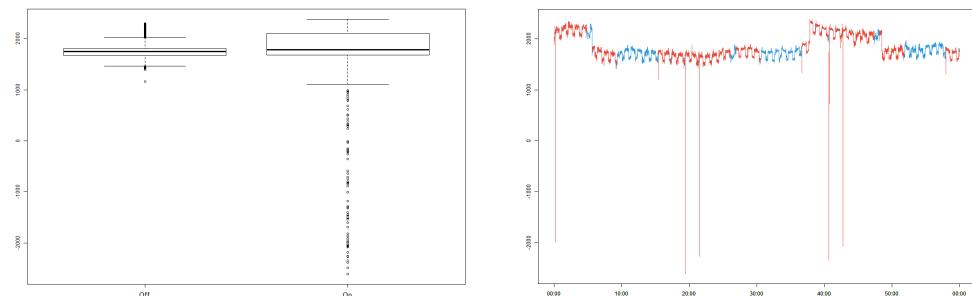
Nos interesan valores altos de la información mutua, ya que esto significa que conocer el valor de una variable predictora reduce la incertidumbre de nuestras variables clase. Basándonos en los resultados obtenidos en la Tabla 4.1, se observa claramente que las variables de intensidad y voltaje, en sus tres fases, son completamente irrelevantes para conocer el valor de las variables clase y por tanto para la construcción del modelo de clasificación. En la Figura 4.4 se justifican estos resultados mediante un estudio gráfico más profundo. Para cada variable de consumo, solamente en su primera fase para no entorpecer la lectura, se muestra en la parte izquierda un diagrama de caja de sus valores con respecto a cada posible valor de la variable clase Spindle (para las otras dos clases se obtienen resultados similares). En la parte derecha se muestran las evoluciones de cada variable predictora a lo largo del intervalo de una hora que dura el conjunto de datos, coloreadas en diferentes tonos en base a los distintos valores de la variable clase (en color azul, el *spindle* está apagado; en color rojo, encendido).



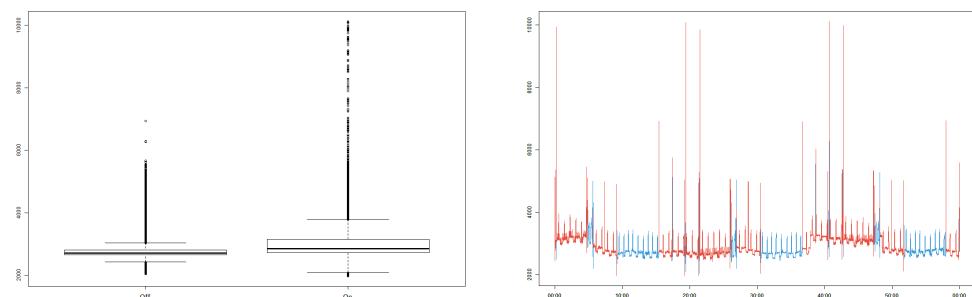
(a) Variable predictora Ia.A.



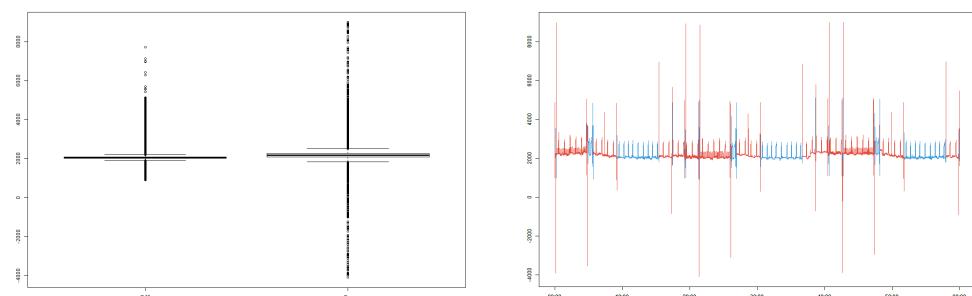
(b) Variable predictora Va.V.



(c) Variable predictora Qa.var.



(d) Variable predictora Sa.VA.



(e) Variable predictora Pa.W.

Figura 4.4: Estudio de las relaciones entre las variables predictoras del problema de eficiencia energética con la variable clase Spindle para la aproximación multi-etiqueta.

En estas gráficas se refuerzan los resultados obtenidos en la selección de variables. Para las variables de intensidad y voltaje, sus diagramas de caja son exactamente iguales para ambos valores de la variable clase, i.e., no hay ninguna distinción significativa en la distribución de estas predictoras cuando la variable Spindle toma distintos valores, por lo que se concluye que estas variables predictoras son irrelevantes para la clasificación de las variables clase. A modo de comparación, en las otras tres variables de consumo se observa cómo sus diagramas de caja difieren significativamente para los distintos valores de la variable clase. Los mismos comentarios se aplican sobre las gráficas que muestran la evolución temporal de las variables predictoras.

Por último, en la Tabla 4.1 se observa que entre las distintas variables clase también existe una información mutua no significativa. Este hecho queda justificado si atendemos a la distribución de los datos de ejemplo en base a las posibles configuraciones de las variables clase (Figura 4.2). Se observa que no hay una distribución uniforme de los datos entre las configuraciones, es más, está lejos de acercarse a ésta. Por ejemplo, se observa que hay dos configuraciones predominantes con casi el 70 % de los ejemplos entre ambas, o también que el servo del eje Y está encendido sólo si el *spindle* también lo está. Es interesante de esta manera construir un modelo de clasificación que tenga en cuenta estas relaciones entre las variables clase.

Tabla 4.2: Distribución de los datos entre las diferentes configuraciones de clases para el problema de eficiencia energética en su aproximación multi-etiqueta.

Spindle	Eje X	Eje Y	Porcentaje
Off	Off	Off	7,3 %
Off	Off	On	0 %
Off	On	Off	29,0 %
Off	On	On	0 %
On	Off	Off	39,7 %
On	Off	On	2,3 %
On	On	Off	13,4 %
On	On	On	8,3 %

Las nueve variables predictoras escogidas tras la selección de variables, i.e., las tres fases de las variables de consumo de potencia activa, reactiva y aparente, se han utilizado para construir tres modelos de clasificación. El primero no tiene en cuenta las relaciones entre las variables clase, ya que construye una red Bayesiana para cada una de ellas de manera independiente siguiendo la estrategia de *Binary Relevance* (BR). Los otros dos sí tienen en cuenta las posibles relaciones, siendo estos el modelo investigado MBC, aprendido de manera *wrapper* como en [Bielza et al. \[2011\]](#), y nuestro modelo propuesto MBCTree, guiado también por una estrategia *wrapper* como se detalla

Tabla 4.3: Rendimiento en términos de precisión predictiva de los modelos de clasificación entrenados para el problema de eficiencia energética en su aproximación multi-etiqueta.

	Mean accuracy (ec. 2.4)			Global accuracy (ec. 2.3)
	Spindle	Eje X	Eje Y	
BR (red Bayesiana)	0.8646	0.6899	0.9284	0.5813
MBC	0.8036	0.6964	0.9332	0.5977
MBCTree	0.8572	0.7336	0.9520	<b>0.6340</b>

en el Algoritmo 1. Para construir los modelos ha sido necesario primero discretizar las variables predictoras, lo que se ha realizado de manera no supervisada en cinco intervalos de igual número de muestras. De esta manera, cada modelo se ha aprendido desde los datos del primer ciclo completo de la máquina y se ha evaluado sobre el segundo ciclo completo, conduciendo a los resultados que se muestran en la Tabla 4.3.

En estos resultados obtenidos se observa en primer lugar que las técnicas de aprendizaje automático están extrayendo conocimiento satisfactoriamente desde el conjunto de datos. El *accuracy* sobre cada variable clase es mayor para los tres clasificadores que el valor de referencia comentado anteriormente, i.e., los modelos no aprenden a decir el valor de la clase con mayor frecuencia. Si atendemos a la medida *global accuracy*, la cual considera un acierto cuando se predice correctamente el valor de las tres variables clase a la vez, el modelo que no tiene en cuenta las dependencias entre las clases está obteniendo peores resultados. De entre los otros dos que sí modelan estas dependencias, nuestro modelo MBCTree es el que mayor rendimiento ha obtenido.

### 4.3. Aproximación multi-dimensional

En una segunda aproximación al problema, el objetivo es predecir el grado de consumo en el que se encuentra cada elemento activo de la máquina en un momento dado. Para ello, el estado anterior de encendido se divide ahora en tres estados adicionales: consumo mínimo (color verde), consumo intermedio (color amarillo) y consumo alto (color rojo). En la Figura 4.5 se detalla el porcentaje de ejemplos del conjunto de datos etiquetados como cada posible estado de los tres elementos activos.

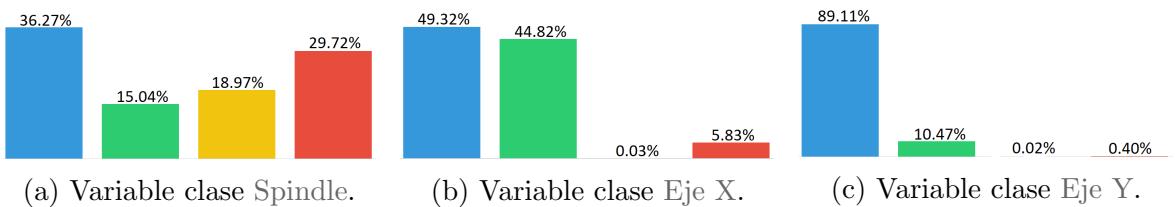


Figura 4.5: Distribución del conjunto de datos del problema de eficiencia energética en su versión multi-dimensional entre cada posible valor de las variables clase.

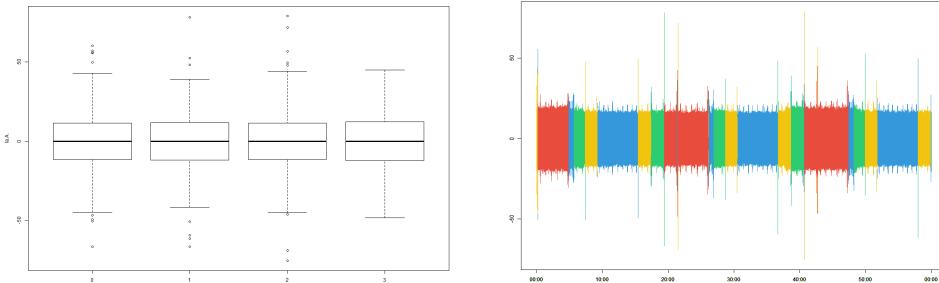
Se puede observar que las variables clase asociadas con los dos servos en el eje X e Y no han sufrido una variación significativa respecto a la configuración multi-etiqueta, ya que casi siempre que estos elementos están encendidos es porque están en un grado de consumo mínimo. En cambio, los instantes de encendido de la variable clase Spindle se han distribuido más equitativamente entre los tres posibles grados de consumo.

El proceso de selección de variables basado en la información mutua (Tabla 4.4) deriva las mismas conclusiones que en el problema multi-etiqueta: las variables de consumo de corriente y voltaje son irrelevantes para inducir conocimiento sobre las variables clase de nuestro problema. Por otro lado, se observa que la información mutua entre cada par de variables ha aumentado con respecto a la aproximación multi-etiqueta, lo que se debe simplemente a que cada variable clase tiene más posibles valores. El ratio de información mutua divide esta medida entre la entropía de la primera variable para no favorecer aquellas con más posibles valores. Si bien, no nos afecta que ocurra esta diferencia respecto a la versión multi-etiqueta.

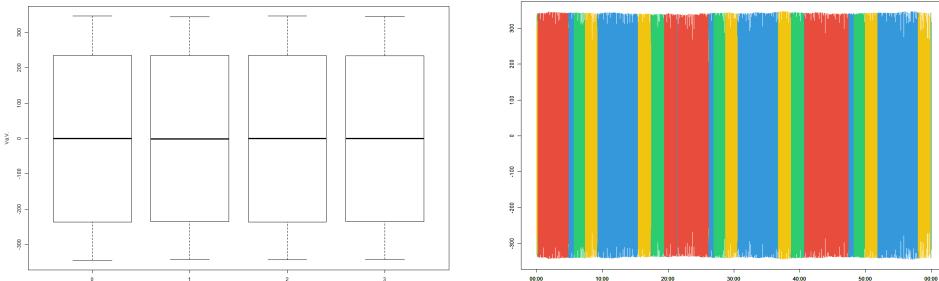
Tabla 4.4: Información mutua entre todas las variables (filas) del problema de eficiencia energética y las tres variables clase (columnas) para la aproximación multi-dimensional.

	Spindle	Eje X	Eje Y
Ia.A	<b>0.004</b>	<b>0.040</b>	<b>0.004</b>
Ib.A	<b>0.044</b>	<b>0.049</b>	<b>0.004</b>
Ic.A	<b>0.044</b>	<b>0.051</b>	<b>0.005</b>
Pa.W	0.617	0.273	0.173
Pb.W	0.702	0.316	0.203
Pc.W	0.640	0.297	0.194
Qa.var	0.372	0.361	<b>0.055</b>
Qb.var	0.393	0.355	<b>0.040</b>
Qc.var	0.464	0.416	<b>0.060</b>
Sa.VA	0.468	0.289	0.153
Sb.VA	0.474	0.284	0.117
Sc.VA	0.456	0.313	0.118
Va.V	<b>0.001</b>	<b>0.002</b>	<b>0.001</b>
Vb.V	<b>0.001</b>	<b>0.002</b>	<b>0.001</b>
Vc.V	<b>0.001</b>	<b>0.001</b>	<b>0.001</b>
Spindle	-	0.470	0.351
Eje X	0.470	-	<b>0.047</b>
Eje Y	0.351	<b>0.047</b>	-

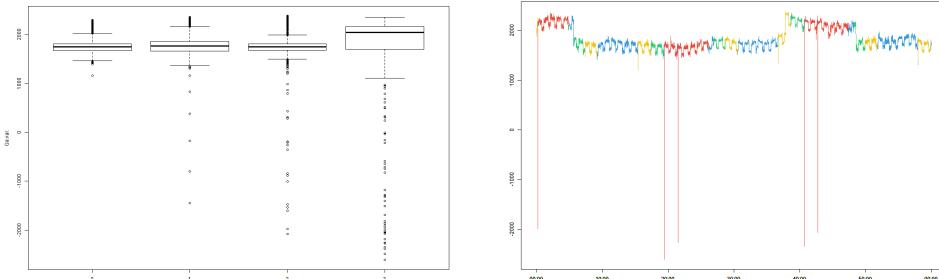
En la Figura 4.6 se justifican estos resultados de igual manera que ocurría en el problema multi-etiqueta. Los diagramas de caja de las variables predictoras irrelevantes son idénticos para los cuatro posibles valores de la clase Spindle.



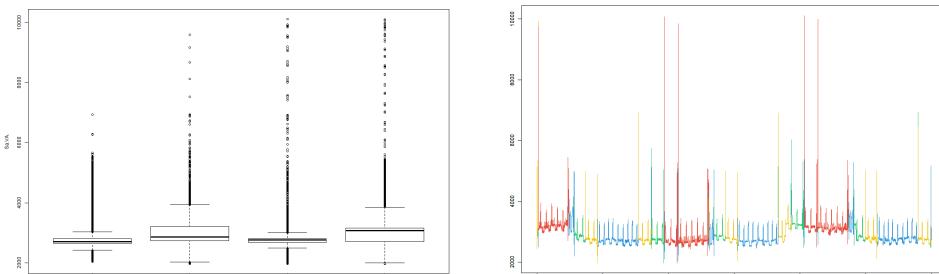
(a) Variable predictora Ia.A.



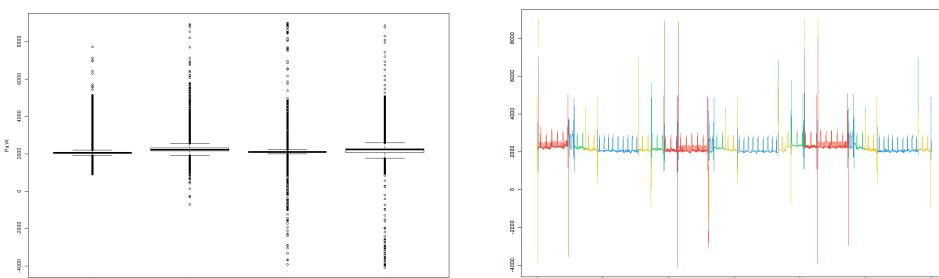
(b) Variable predictora Va.V.



(c) Variable predictora Qa.var.



(d) Variable predictora Sa.VA.



(e) Variable predictora Pa.W.

Figura 4.6: Estudio de las relaciones entre las variables predictoras del problema de eficiencia energética con la variable clase Spindle para la aproximación multi-dimensional.

Finalmente, se ha aplicado el mismo proceso de entrenamiento y evaluación sobre el conjunto de datos etiquetado con los grados de consumo de los elementos de la máquina. Los resultados obtenidos se exponen en la Tabla 4.5. De nuevo, el modelo que no tiene en cuenta las relaciones entre las variables clase es el que peor rendimiento obtiene en términos de precisión predictiva. En este problema multi-dimensional, la diferencia de rendimiento es mucho más evidente dado que las relaciones entre las variables clase son más fuertes al existir un mayor número de configuraciones conjuntas. Nuestro modelo MBCTree también repite como el clasificador que consigue los mejores resultados. Si bien, ha sido necesario aumentar el número mínimo de instancias desde 50 hasta 500 para crear un nodo hoja del árbol. La razón se debe a que al ser más grande el espacio de configuraciones de las variables clase, se necesitan más ejemplos para aprender correctamente los MBC que conforman las hojas del árbol.

Por otro lado, también se puede observar una degradación significativa en el rendimiento sobre la variable clase Spindle con respecto al problema multi-etiqueta, lo que se justifica por la existencia de más posibles valores de la variable y una distribución de los datos equitativa entre cada posible valor. En cambio, aunque las dos variables clase de los servos tienen el mismo número de posibles valores, se cumple que estos elementos están generalmente en un consumo mínimo cuando en el problema multi-etiqueta se simplificaba en un estado de encendido. Es decir, los consumos intermedio y alto aparecen raramente en estas variables. Por ello, se tiene una distribución similar de estas variables frente al problema multi-etiqueta, lo que conduce a la obtención de rendimientos similares. Atendiendo a la medida *global accuracy*, se observa razonadamente una ligera reducción en el rendimiento de los clasificadores entrenados para este problema multi-dimensional más difícil.

Tabla 4.5: Rendimiento en términos de precisión predictiva de los modelos de clasificación entrenados para el problema de eficiencia energética en su aproximación multi-dimensional.

	Mean accuracy (ec. 2.4)			Global accuracy (ec. 2.3)
	Spindle	Eje X	Eje Y	
BR (red Bayesiana)	0.6872	0.7011	0.9071	0.4940
MBC	0.7060	0.7103	0.9162	0.5743
MBCTree	0.7128	0.7186	0.9275	<b>0.5856</b>



# Capítulo 5

## Flujos de datos multi-dimensionales con cambio de concepto

### 5.1. Definición y contexto

Debido a los constantes avances en la transmisión y el almacenamiento de datos, los operadores de redes se enfrentan continuamente a una gran cantidad de eventos en línea que se producen a una velocidad muy alta. Este tipo de información se conoce como flujos de datos (en inglés, *data streams*), el cual se caracteriza por su aspecto de cambio de concepto [Widmer and Kubat, 1996]. Un cambio de concepto se refiere principalmente a un escenario de aprendizaje supervisado en línea cuando la relación entre las variables predictoras y las variables clase cambia con el tiempo. Por ejemplo, en un contexto de Industria 4.0, la monitorización y captura de datos continua de un proceso industrial se corresponde con este escenario de flujos de datos comentado. En el problema anteriormente expuesto, se quería predecir qué máquinas están activas durante un proceso industrial en base a sus huellas energéticas. Si bien, éstas seguramente cambian a lo largo del tiempo debido al desgaste y deterioro de las máquinas, produciéndose el así nombrado cambio de concepto.

Existen múltiples categorizaciones no mutuamente excluyentes de un cambio de concepto, por ejemplo dependiendo de su ratio de cambio como indica Gama and Castillo [2006], o en base a su severidad y frecuencia como detalla Minku et al. [2010]. En la Figura 5.1 se muestra una visualización gráfica de diferentes tipos de cambio de concepto a lo largo del tiempo. Se comprueba que estas categorías no son mutuamente excluyentes. Por ejemplo, el cambio de concepto que se expone como recurrente también es repentino. Un escenario de flujos de datos tiene requisitos adicionales en relación a recursos de memoria (i.e., el flujo no se puede almacenar por completo en memoria), y tiempo (i.e., el flujo se debe procesar continuamente y el modelo de clasificación aprendido tiene que estar preparado en cualquier momento para predecir).

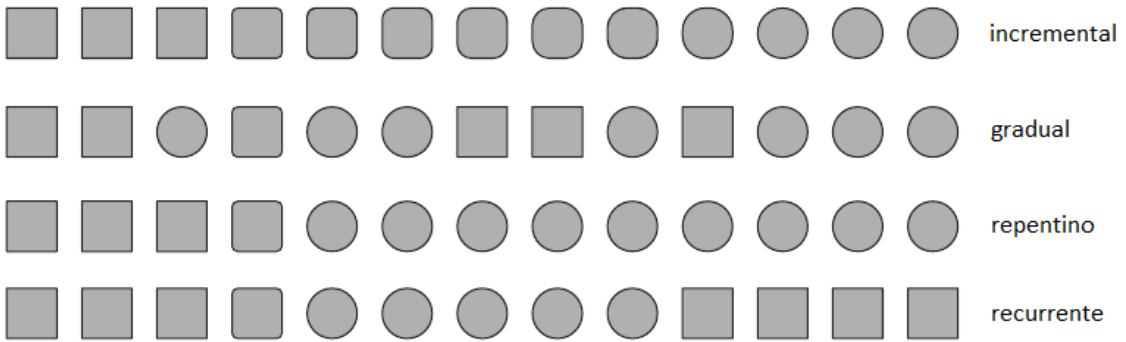


Figura 5.1: Visualización de un cambio de concepto a lo largo del tiempo. En la imagen, el tiempo progresa hacia la derecha. Cada figura corresponde al valor de la variable clase, y por tanto representa el concepto actual. Se observa que el concepto cambia a lo largo del tiempo. Este cambio de concepto se puede categorizar en base a diferentes factores. Imagen extraída de [Krawczyk et al. \[2017\]](#).

El problema de clasificación de flujos de datos ha sido extensamente estudiado en la literatura. Todas las propuestas tienen el objetivo principal de hacer frente al cambio de concepto y mantener el modelo de clasificación actualizado a lo largo de los flujos continuos de datos. Se suelen componer de un método que monitoriza y detecta el cambio de concepto y un método de adaptación para actualizar el modelo de clasificación a lo largo del tiempo. Si bien, la mayoría del trabajo en este ámbito se ha centrado en clasificar flujos de datos con variables clase uni-dimensionales, donde una instancia debe asignarse a una única variable clase. [Gama et al. \[2014\]](#) realizó una revisión del estado del arte en la adaptación al cambio de concepto en este contexto de clasificación uni-dimensional. El problema de clasificar flujos de datos multi-dimensionales permanece en gran parte inexplorado y sólo pocos métodos se han propuesto. A continuación revisamos la literatura y recopilamos los modelos propuestos para clasificar flujos de datos multi-dimensionales (Tabla 5.1).

## 5.2. Aproximaciones en la literatura

Existen dos aproximaciones para hacer frente al posible cambio de concepto que puede ocurrir en los flujos de datos. La primera mantiene un único modelo en activo, el cual se actualiza en el tiempo para representar el concepto más reciente. La segunda técnica y más popular se denomina *ensemble learning*. Un *ensemble* es un conjunto de clasificadores cuyas predicciones se combinan para predecir nuevas instancias. Los clasificadores individuales pueden representar diferentes conceptos de los datos, por lo que la constante actualización del *ensemble* permite modelar posibles cambios de concepto. En [Krawczyk et al. \[2017\]](#) se revisa la literatura más relevante en relación a esta técnica de *ensemble learning*.

### 5.2.1. Basadas en un único modelo

[Borchani et al. \[2016\]](#) propuso un modelo adaptativo, *Locally Adaptive-MB-MBC*, basado en MBCs, el cuál extiende el propio algoritmo estacionario MB-MBC [[Borchani et al., 2012](#)] para hacer frente al aspecto de cambio de concepto de los flujos de datos. A diferencia de la mayoría de los métodos que utilizan para ello la técnica de *ensemble learning*, esta aproximación monitoriza el posible cambio de concepto sobre el tiempo en un único clasificador base (un MBC) utilizando la verosimilitud de los datos más recientes al modelo actual y el test de Page-Hinkley [[Page, 1954](#), [Hinkley, 1971](#)] para decidir cuándo se produce un cambio de concepto. En el caso de que sea detectado, el método adapta el MBC actual de manera local sobre cada nodo que ha cambiado en el nuevo concepto, extendiendo para ello el algoritmo HITON [[Aliferis et al., 2010a,b](#)]. De esta manera, no es necesario aprender la red al completo desde cero. Si bien, también se puede utilizar una adaptación global sobre todo el modelo para representar el nuevo concepto mediante el método que los autores denominan *Global Adaptive-MB-MBC*. Esta aproximación global puede ser interesante en el caso de cambios de concepto abruptos [[Gama and Castillo, 2006](#)] o severos [[Minku et al., 2010](#)]. Se destaca finalmente que estos modelos son los únicos propuestos en la literatura basados en MBCs para resolver el problema de flujos de datos multi-dimensionales. Todos los modelos que se exponen a continuación se basan en un contexto multi-etiqueta.

[Xioufis et al. \[2011\]](#) también utilizan un único modelo para abordar además un problema especial denominado desbalanceo de clases, i.e., la asimetría en la distribución de instancias positivas y negativas para todas o alguna de las etiquetas. Para resolver este problema, los autores proponen un clasificador de múltiples ventanas que mantiene dos ventanas de tamaño fijo para cada etiqueta: una para las instancias positivas y otra para las negativas. La idea de esta ventana parametrizada es tomar más o menos muestras de las instancias con menor y mayor frecuencia, respectivamente, para mantener un ratio deseado en la distribución de las clases. Los autores asumen un cambio de concepto independiente para cada etiqueta y adoptan un modelo conocido como relevancia binaria, i.e., descomponen el problema de clasificación multi-etiqueta en varios problemas binarios, uno por cada etiqueta, por lo que no se modela las interacciones entre ellas. Establecen el modelo *kNN* como clasificador base debido a su naturaleza incremental, i.e., no es necesario actualizar ningún modelo conforme se introducen nuevos ejemplos en las ventanas. El método propuesto también se podría acoplar con un clasificador base no incremental, aunque esto requeriría reconstruir el modelo de clasificación en períodos regulares de tiempo con su consecuente reducción en la habilidad de adaptarse rápidamente ante cambios de concepto repentinos. De esta

manera, el método propuesto no incorpora ningún método de detección del cambio de concepto, sino que constantemente actualiza las ventanas positivas y negativas de cada etiqueta mediante la inclusión de nuevas instancias y eliminación de las más antiguas, siendo así capaz de modelar el concepto actual.

### 5.2.2. Basadas en *ensemble*

[Wei et al. \[2009\]](#) resuelven el problema de clasificación multi-etiqueta mediante el método de relevancia binaria, el cual extienden en su trabajo posterior [\[Qu et al., 2009\]](#) con el método mejorado de relevancia binaria propuesto en [Godbole and Sarawagi \[2004\]](#), en el que las salidas de un modelo sobre una etiqueta se usan en el espacio de predictoras del modelo sobre la siguiente etiqueta. Esta forma de apilado permite tener en cuenta las relaciones entre las etiquetas. Aunque los autores no utilizan ningún método de detección del cambio de concepto, incorporan la técnica de *ensemble learning* para resolverlo, tal que el *ensemble* con un número fijo de clasificadores base se actualiza continuamente sobre el tiempo mediante la adición de nuevos clasificadores entrenados sobre los bloques de datos más recientes y la eliminación de los clasificadores más antiguos. Se sigue una estrategia de votación por mayoría ponderada para clasificar nuevas instancias, tal que el peso de cada clasificador individual se calcula dinámicamente en base a su rendimiento en los  $k$  vecinos más cercanos a la instancia que se quiere predecir en el bloque de datos más reciente. De esta manera, se consigue que aquellos clasificadores que más se asimilen al concepto actual tengan un mayor peso en la predicción final. Los autores utilizan naive Bayes, el árbol de clasificación C4.5 y máquinas de vectores de soporte (SVM, del inglés *support vector machines*) como clasificadores base para evaluar este método propuesto.

[Trajdos and Kurzynski \[2015\]](#) propusieron una extensión basada en flujos de datos del modelo de relevancia binaria utilizando una matriz de confusión borrosa para corregir las decisiones de los clasificadores base en el *ensemble*. Este método evalúa el clasificador multi-etiqueta base y después calcula medidas de competencia cruzada y específica en una etiqueta para ajustar su predicción final. Aunque no se utiliza una técnica de detección del cambio de concepto, la corrección del modelo se actualiza conforme progresan el flujo de datos utilizando una ventana deslizante. Los autores evaluaron el método con naive Bayes como clasificador base.

Otro método basado en *ensembles* para la clasificación de flujos de datos multi-etiqueta se propuso en [Kong and Philip \[2011\]](#). Se basa en utilizar un *ensemble* de múltiples árboles de decisión aleatorios (*random trees*) propuestos por [Zhang et al. \[2010\]](#), donde los nodos del árbol se construyen por medio de variables y valores de partición aleatorios. Para solucionar el cambio de concepto, los autores utilizan lo que

denominan funciones debilitadoras (*fading functions*) en los nodos de cada árbol con el fin de debilitar la influencia de datos pasados, en vez de utilizar un mecanismo de detección del punto de cambio de concepto. Estas funciones incluyen un parámetro  $\lambda$ , llamado factor debilitador, el cual indica la velocidad de los efectos debilitadores.

En [Read et al. \[2012\]](#) se propuso el primer modelo de clasificación de flujos de datos multi-etiqueta con un esquema incremental por instancias y se discute su mejor adecuación frente a las propuestas incrementales por bloques debido a su habilidad para aprender cada ejemplo tal como llega, manteniendo de esta manera un clasificador relevante. El método propuesto extiende el clasificador uni-dimensional *Hoeffding tree* de [Domingos and Hulten \[2000\]](#) mediante la utilización de una definición multi-etiqueta de entropía y conjuntos podados multi-etiqueta de entrenamiento en cada nodo hoja del árbol. Para dar solución al cambio de concepto, los autores usan como detector ADWIN *Bagging* de [Bifet et al. \[2009\]](#), el cual consiste en el método *bagging* de [Oza \[2005\]](#) extendido con una ventana deslizante adaptativa. Cuando se detecta el cambio de concepto, el clasificador con peor rendimiento se reemplaza por uno nuevo. Los autores también extienden los métodos estacionarios de relevancia binaria, *ensembles* de relevancia binaria [[Read et al., 2011](#)], *ensembles* de conjuntos podados [[Read et al., 2008](#)] y *Hoeffding trees* multi-etiqueta [[Clare and King, 2001](#)], mediante la inclusión del algoritmo ADWIN para detectar los potenciales cambios de concepto.

[Shi et al. \[2014a\]](#) propusieron un método eficiente y efectivo para detectar cambios de concepto en el que las etiquetas se agrupan utilizando reglas de asociación, teniendo en cuenta de esta manera las dependencias entre ellas. Más tarde, [Shi et al. \[2014b\]](#) diseñaron un algoritmo de aprendizaje incremental por clases, el cual reconoce de manera dinámica nuevas combinaciones frecuentes de etiquetas. Este método extiende la anterior propuesta de [Read et al. \[2012\]](#) con esta estrategia de aprendizaje.

Los problemas relacionados con los costes de etiquetado de flujos de datos multi-etiqueta se discutieron en [Wang et al. \[2012\]](#), además de contribuir al problema de distribución de clases no balanceadas. Los autores derivan una función de pérdida teórica para el *ensemble* propuesto y un aprendizaje activo para seleccionar los ejemplos que minimicen dicha función. Esto permite utilizar menos instancias etiquetadas para entrenar y detectar un cambio de concepto de acuerdo al etiquetado de los ejemplos más inciertos. Para solucionar el problema del cambio de concepto, los autores usan un esquema de ponderación para actualizar continuamente los pesos del *ensemble* en base a la precisión predictiva de los clasificadores base.

[Song and Ye \[2014\]](#) propusieron un *ensemble* dinámico multi-etiqueta. Este método utiliza clasificadores multi-etiqueta basados en agrupaciones (MLCC, del inglés *multi-label cluster-based classifiers*) que se combinan mediante una estrategia de

votación que utiliza dos pesos basados en la precisión sobre el bloque de datos actual y la similaridad entre bloques. En comparación con las aproximaciones existentes en la literatura con una ventana de tamaño fijo, los autores proponen un *ensemble* dinámico en el que el número de clasificadores base cambia de acuerdo a si hay cambio de concepto o no. Si no lo hay, el número de clasificadores se incrementa para obtener una mejor precisión. En caso contrario, el número de clasificadores se reduce automáticamente.

Finalmente, Wang et al. [2017] propusieron un *ensemble* basado en ML- $k$ NN [Zhang and Zhou, 2007]. Los autores propusieron una función sobre tres factores para combinar de manera eficiente las predicciones de los clasificadores en el *ensemble*: (1) la confianza sobre cada etiqueta, (2) la diferencia en tiempo y (3) la diferencia de distancia.

Tabla 5.1: Recopilación de los métodos de clasificación de flujos de datos multi-dimensionales. Extendido de Borchani et al. [2016].

Método y referencia	Clasificador base	Estrategia de adaptación
<i>Ensemble</i> de relevancia binaria [Qu et al., 2009] (MBR)	Naive Bayes, C4.5, SVM	<i>Ensemble</i> evolutivo. Sin detección
Clasificador de múltiples ventanas [Xioufis et al., 2011] (MWC)	$k$ NN	Dos ventanas de tamaño fijo. Sin detección
<i>Streaming multi-label random trees</i> [Kong and Philip, 2011] (SMART)	Random tree	Función <i>fading</i> . Sin detección
<i>Ensemble</i> con costes de etiquetado [Wang et al., 2012] (LBEF)	-	<i>Ensemble</i> evolutivo. Sin detección
<i>Ensemble</i> de Hoeffding trees multi-etiqueta con conjuntos podados en las hojas [Read et al., 2012] (EaHT <sub>PS</sub> )	Hoeffding tree	<i>Ensemble</i> evolutivo. Detección: algoritmo ADWIN
EaHT <sub>PS</sub> incremental por clases [Shi et al., 2014b] (EaHT <sub>CL</sub> )	Hoeffding tree	<i>Ensemble</i> evolutivo. Detección: algoritmo ADWIN
<i>Ensemble</i> dinámico multi-etiqueta [Song and Ye, 2014] (MLDE)	MLCC	<i>Ensemble</i> evolutivo. Sin detección
<i>Ensemble</i> de relevancia binaria con una matriz de confusión borrosa [Trajdos and Kurzynski, 2015] (FCM)	Naive Bayes	Actualizar corrección. Sin detección
<i>Locally Adaptive-MB-MBC</i> [Borchani et al., 2016] (LA-MB-MBC)	MBC	Adaptación local del MBC. Detección: Page-Hinkley test
<i>Ensemble</i> ponderado sobre ML- $k$ NN [Wang et al., 2017] (SWMEC)	ML- $k$ NN	<i>Ensemble</i> evolutivo. Sin detección

# Capítulo 6

## Cierre del documento

En este capítulo se extraen las principales conclusiones del trabajo realizado y posteriormente se continua con una breve discusión sobre las mismas. Asimismo, se proponen algunas líneas interesantes de trabajo futuro. Finalmente, se proporciona una evaluación personal del autor.

### 6.1. Conclusiones

Las principales conclusiones alcanzadas sobre el trabajo de investigación realizado y su aplicación al problema de eficiencia energética, identificadas con una letra por cada capítulo presentado en el documento, han sido:

- **A1:** Hemos revisado la literatura existente sobre los MBCs y hemos observado que se trata de una línea de investigación actual que sigue presentando nuevos trabajos y aplicaciones.
- **A2:** Hemos comprobado la complejidad súper exponencial del tamaño del espacio de estructuras de los MBCs.
- **A3:** El conjunto de medidas existentes en la literatura para evaluar el rendimiento de clasificadores multi-dimensionales es escaso, por lo que hemos aportado dos nuevas medidas apropiadas para su evaluación.
- **B1:** Hemos desarrollado un nuevo clasificador multi-dimensional, MBCTree, que ha conseguido mejores resultados en términos de precisión predictiva frente a los MBCs en un estudio experimental sobre datos sintéticos y sobre el problema de eficiencia energética.
- **B2:** La mejora en predicción del MBCTree con respecto a los MBCs se consigue a expensas de añadir complejidad al entrenamiento del modelo.
- **C1:** Hemos explorado la aproximación mediante técnicas de aprendizaje automático al problema de eficiencia energética en la Industria 4.0, obteniendo resultados prometedores que motivan continuar con un estudio más profundo.

- **C2:** Tener en cuenta los diferentes grados de consumo de los elementos activos de la máquina no empeora en gran medida los resultados, lo cual genera interés en realizar una investigación más exhaustiva de esta aproximación.
- **D1:** Aplicar un método que resuelva el problema de eficiencia energética en una línea de producción en tiempo real requerirá hacer frente a los potenciales cambios de concepto asociados con los cambios en las huellas energéticas de los elementos de la máquina, originados por el desgaste y deterioro con el tiempo de los mismos.
- **D2:** No existe mucha literatura para la clasificación de flujos de datos multi-dimensionales con presencia de cambios de concepto a lo largo del tiempo. La mayoría de las propuestas se basan en un modelo basado en *ensemble*.

## 6.2. Discusión y trabajo futuro

En primer lugar, como foco de investigación del proyecto es interesante la conclusión **A1** respecto a los MBCs. Junto con este avance, la falta de una revisión exhaustiva sobre el trabajo existente en la literatura de estos clasificadores nos ha hecho trabajar en la redacción de un artículo que pretendemos publicar en la revista de revisión por pares ACM Computing Surveys. Además, sería interesante aplicar estos modelos de clasificación multi-dimensional a una mayor gama de dominios. La aplicación de eficiencia energética en la Industria 4.0 desarrollada en este proyecto es un ejemplo innovador de otro dominio sobre el que se ha aplicado un MBC satisfactoriamente, y que se debe añadir al listado de aplicaciones recopilado en la Sección 2.5.

Las redes Bayesianas dinámicas [Dean and Kanazawa, 1989] y de tiempo continuo [Nodelman et al., 2002] permiten razonar sobre sistemas que evolucionan con el tiempo. Las primeras dividen el tiempo en instantes separados por intervalos de igual duración, mientras que las últimas evitan usar esta granularidad del tiempo al basarse en procesos homogéneos de Markov [Norris, 1998]. Stella and Amer [2012] extendieron los clasificadores de redes Bayesianas de tiempo continuo para clasificar una variable estática dada una evidencia en tiempo continuo. Si bien, en la literatura no se ha planteado la extensión de estos modelos al contexto multi-dimensional, por lo que sería interesante investigar y proponer las extensiones de las redes Bayesianas dinámicas y de tiempo continuo para el problema de clasificación multi-dimensional.

El nuevo método MBCTree desarrollado no sólo ha conseguido mejores resultados que los MBCs en un estudio sobre datos sintéticos como se concluye en **B1**, sino también sobre la aplicación real de eficiencia energética. Si bien, es interesante plantear un estudio futuro más profundo en el que se comparen ambos métodos sobre conjuntos de datos con diferente dimensionalidad de variables y clases, así como poder controlar el

grado de diferencia entre las distribuciones de los MBCs hoja del MBCTree generado inicialmente sobre el que se simulan los datos. De esta manera, se podrá concluir en las características de aquellos problemas en los que un MBCTree es más apropiado.

En **A2** se concluye la necesidad de algoritmos eficientes que se muevan por el espacio de estructuras de los MBCs dada la complejidad súper exponencial de éste. Sería interesante plantear la misma complejidad del espacio de estructuras de los MBCTrees, aunque se conoce ya que es mucho mayor que aquella de los MBCs y por ello obteniendo los resultados que concluyen en **B2**, por la razón obvia de que son múltiples los MBCs que hay considerar en las hojas y a diferentes profundidades del árbol frente a un único modelo MBC. Motivados por ello, queremos investigar tres estrategias:

- Un algoritmo de aprendizaje *filter* frente al propuesto *wrapper* que permita reducir la carga computacional del entrenamiento del modelo. La verosimilitud de los datos a una estructura MBCTree se propone de manera sencilla como el producto de las verosimilitudes de los subconjuntos de datos que alcanzan cada MBC hoja del árbol. Si bien, esto resultaría en modelos sobre-ajustados a los datos de entrenamiento por lo que es necesaria la propuesta de una verosimilitud penalizada, como la puntuación BIC, que se ajuste a una estructura de MBCTree.
- Una selección aleatoria de los nodos internos del MBCTree junto con una estrategia de *ensemble*, siguiendo una idea similar al método *random forest*. El algoritmo de aprendizaje *wrapper* propuesto tiene el punto crítico de carga computacional en el proceso de selección de la mejor variable predictora para ser nodo interno del árbol, dado que para ello hay que calcular un MBC para cada valor de todas las predictoras. Si esta selección se aleatoriza y se compensa con una estrategia de *ensemble*, este punto crítico desaparece. Para abordar este planteamiento será necesario pensar una forma de consenso multi-dimensional en la fase de predicción entre todos los MBCTrees que conformen el *ensemble*.
- El algoritmo de aprendizaje propuesto se puede paralelizar, dado que el aprendizaje de cada subárbol una vez se ha seleccionado la mejor variable como nodo interno son procesos independientes. De igual forma, el proceso de selección de la mejor variable como nodo interno también es paralelizable en la forma que se puede evaluar el rendimiento de cada variable predictora como nodo interno de manera independiente.

Por otro lado, es de interés la extensión de métodos que permitan trabajar con variables predictoras continuas en ambos modelos MBCs y MBCTrees, sin tener que realizar una discretización ni asumir una distribución Gaussiana de las mismas. Para ello, se pueden investigar la extensión de técnicas basadas en *kernels* para los MBCs, así como la inclusión de diferentes técnicas de discretización a la hora de seleccionar la mejor variable como nodo interno del MBCTree.

Respecto a la aplicación de eficiencia energética en la Industria 4.0, se concluye en **C1** y **C2** con la obtención de unos primeros buenos resultados, pero con la necesidad de seguir profundizando sobre el problema. El entrenamiento y evaluación de los modelos sobre una mayor cantidad de ciclos de producción resulta de interés, así como mantener una colaboración más cercana con la empresa que permita concluir en las variables predictoras que más información proporcionen.

Es interesante también investigar el problema de eficiencia energética con grados de consumo como un problema de clasificación multi-dimensional bajo una función de pérdida sensible a los costes entre los diferentes grados de consumo, i.e., que predecir que un elemento está a máxima potencia cuando en verdad no está consumiendo tenga un mayor coste que se prediga a máxima potencia y en verdad esté a una intermedia.

Finalmente, se concluye en **D1** la necesidad de utilizar técnicas adaptativas para hacer frente al potencial cambio de concepto en relación al desgaste energético de las máquinas industriales. Es atractiva la extensión de los MBCTrees a este contexto de flujos de datos, posiblemente mediante técnicas de *ensemble* como se concluye en **D2**.

### 6.3. Evaluación personal

Este proyecto ha impregnado en mí la esencia al completo de un trabajo de investigación. En un primer momento, me ha resultado bastante difícil la búsqueda y lectura de toda la literatura relevante, pero conforme vas atando hilos entre los distintos trabajos te das cuenta de que es un proceso muy agradecido. El hecho de no solamente revisar el estado del arte, sino también proponer un nuevo método y abrir una nueva línea de investigación al mundo científico me ha parecido muy bonito. Al mismo tiempo, la escritura de un artículo científico me ha resultado muy dura dado que se trata del primero que realizo en esta nueva etapa como investigador. La ayuda que me han ofrecido mis directores en este aspecto ha sido de mucha importancia.

La colaboración con la empresa Etxe-Tar me ha parecido muy motivadora. Aquellos proyectos que agrupan ambos mundos académico y profesional los considero muy meritorios, y si además se pretende abordar un problema tan innovador y de alta complejidad como el que nos ha ofrecido Etxe-Tar, entonces la motivación que me causa se duplica. En adición, el trabajo realizado se trata de un proyecto completamente nuevo tanto para la empresa como para el grupo de investigación, por lo que participar en la cimentación del mismo ha sido laborioso y muy agradable. Trabajar de la mano de los directores de innovación de esta empresa y participar en primera línea en las diferentes reuniones celebradas ha fortalecido mi experiencia profesional. También me ha resultado muy satisfactorio ver cómo los resultados presentados a la empresa les han sido de utilidad. Como evaluación final, el proyecto ha sido una experiencia perfecta.

# Capítulo 7

## Bibliografía

- A. M. Abdelbar and S. M. Hedetniemi. Approximating MAPs for belief networks is NP-hard and other theorems. *Artificial Intelligence*, 102(1):21–38, 1998.
- C. F. Aliferis, A. Statnikov, I. Tsamardinos, S. Mani, and X. D. Koutsoukos. Local causal and Markov blanket induction for causal discovery and feature selection for classification Part I: Algorithms and empirical evaluation. *Journal of Machine Learning Research*, 11:171–234, 2010a.
- C. F. Aliferis, A. Statnikov, I. Tsamardinos, S. Mani, and X. D. Koutsoukos. Local causal and Markov blanket induction for causal discovery and feature selection for classification Part II: Analysis and extensions. *Journal of Machine Learning Research*, 11:235–284, 2010b.
- A. Antonucci, G. Corani, D. Mauá, and S. Gabaglio. An ensemble of Bayesian networks for multilabel classification. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence*, volume 13, pages 1220–1225, 2013.
- S. Arnborg, D. G. Corneil, and A. Proskurowski. Complexity of finding embeddings in a k-tree. *SIAM Journal on Algebraic Discrete Methods*, 8(2):277–284, 1987.
- M. Benjumeda, C. Bielza, and P. Larrañaga. Tractability of most probable explanations in multidimensional Bayesian network classifiers. *International Journal of Approximate Reasoning*, 93:74–87, 2018.
- C. Bielza and P. Larrañaga. Discrete Bayesian network classifiers: A survey. *ACM Computing Surveys*, 47(1):5, 2014.
- C. Bielza, G. Li, and P. Larrañaga. Multi-dimensional classification with Bayesian networks. *International Journal of Approximate Reasoning*, 52(6):705–727, 2011.

- J. Biesiada and W. Duch. Feature selection for high-dimensional data: A Kolmogorov-Smirnov correlation-based filter. In *Computer Recognition Systems*, pages 95–103. Springer, 2005.
- A. Bifet, G. Holmes, B. Pfahringer, R. Kirkby, and R. Gavaldà. New ensemble methods for evolving data streams. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 139–148. ACM, 2009.
- R. Blanco, I. Inza, M. Merino, J. Quiroga, and P. Larrañaga. Feature selection in Bayesian classifiers for the prognosis of survival of cirrhotic patients treated with TIPS. *Journal of Biomedical Informatics*, 38(5):376–388, 2005.
- H. Borchani, C. Bielza, and P. Larrañaga. Learning CB-decomposable multi-dimensional Bayesian network classifiers. In *Proceedings of the 5th European Workshop on Probabilistic Graphical Models*, pages 25–32, 2010.
- H. Borchani, C. Bielza, P. Martínez-Martí, and P. Larrañaga. Markov blanket-based approach for learning multi-dimensional Bayesian network classifiers: An application to predict the European quality of life-5 dimensions (EQ-5D) from the 39-item Parkinson’s disease questionnaire (PDQ-39). *Journal of Biomedical Informatics*, 45(6):1175–1184, 2012.
- H. Borchani, C. Bielza, C. Toro, and P. Larrañaga. Predicting human immunodeficiency virus inhibitors using multi-dimensional Bayesian network classifiers. *Artificial Intelligence in Medicine*, 57(3):219–229, 2013.
- H. Borchani, P. Larrañaga, J. Gama, and C. Bielza. Mining multi-dimensional concept-drifting data streams using Bayesian network classifiers. *Intelligent Data Analysis*, 20(2):257–280, 2016.
- R. R. Bouckaert. Optimizing causal orderings for generating DAGs from data. In *Proceedings of the 8th Conference on Uncertainty in Artificial Intelligence*, pages 9–16. Elsevier, 1992.
- L. Breiman, J. H. Friedman, R. Olshen, and C. J. Stone. *Classification and Regression Trees*. Wadsworth, 1984.
- G. W. Brier. Verification of forecasts expressed in terms of probability. *Monthey Weather Review*, 78(1):1–3, 1950.

- W. Buntine. Theory refinement on Bayesian networks. In *Proceedings of the 7th Conference on Uncertainty in Artificial Intelligence*, pages 52–60. Morgan Kaufmann Publishers Inc., 1991.
- A. Carbonari, M. Vaccarini, and A. Giretti. Bayesian networks for supporting model based predictive control of smart buildings. In *Dynamic Programming and Bayesian Inference, Concepts and Applications*, pages 3–38. InTech, 2014.
- I. M. Chakravarty, J. D. Roy, and R. G. Laha. Handbook of methods of applied statistics. In *Handbook of Methods of Applied Statistics*. John Wiley & Sons, 1967.
- C. Chow and C. Liu. Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, 14(3):462–467, 1968.
- Y. J. Chu and T. H. Liu. On the shortest arborescence of a directed graph. *Scientia Sinica*, 14:1396–1400, 1965.
- A. Clare and R. King. Knowledge discovery in multi-label phenotype data. In *Principles of Data Mining and Knowledge Discovery*, pages 42–53. Springer, 2001.
- G. F. Cooper and E. Herskovits. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9(4):309–347, 1992.
- G. Corani, A. Antonucci, D. D. Mauá, and S. Gabaglio. Trading off speed and accuracy in multilabel classification. In *European Workshop on Probabilistic Graphical Models*, pages 145–159. Springer, 2014.
- A. P. Dawid. Applications of a general propagation algorithm for probabilistic expert systems. *Statistics and Computing*, 2(1):25–36, 1992.
- P. R. de Waal and L. C. van der Gaag. Inference and learning in multi-dimensional Bayesian network classifiers. In *Proceedings of the 9th European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty, Lecture Notes in Artificial Intelligence*, volume 4724, pages 501–511. Springer, 2007.
- T. Dean and K. Kanazawa. A model for reasoning about persistence and causation. *Computational Intelligence*, 5(2):142–150, 1989.
- K. Deb, S. Agrawal, A. Pratap, and T. Meyarivan. A fast elitist non-dominated sorting genetic algorithm for multi-objective optimization: NSGA-II. In *Proceedings of the 6th International Conference on Parallel Problem Solving from Nature*, pages 849–858. Springer, 2000.

- R. Dechter. Bucket elimination: A unifying framework for reasoning. *Artificial Intelligence*, 113(1-2):41–85, 1999.
- R. Dechter and I. Rish. A scheme for approximating probabilistic inference. In *Proceedings of the 13th Conference on Uncertainty in Artificial Intelligence*, pages 132–141. Morgan Kaufmann Publishers Inc., 1997.
- A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–38, 1977.
- P. Domingos and G. Hulten. Mining high-speed data streams. In *Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 71–80. ACM, 2000.
- U. Fayyad and K. Irani. Multi-interval discretization of continuous-valued attributes for classification learning. *Proceedings of the 13th International Joint Conference on Artificial Intelligence*, 2:1022–1027, 1993.
- J. A. Fernandes, J. A. Lozano, I. Inza, X. Irigoién, A. Pérez, and J. D. Rodríguez. Supervised pre-processing approaches in multiple class variables classification for fish recruitment forecasting. *Environmental Modelling & Software*, 40:245–254, 2013.
- P. Fernandez-Gonzalez, C. Bielza, and P. Larrañaga. Multidimensional classifiers for neuroanatomical data. In *ICML Workshop on Statistics, Machine Learning and Neuroscience (StamLins 2015)*, 2015.
- E. Frank, M. A. Hall, and I. H. Witten. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 2006. The WEKA workbench. Online appendix.
- N. Friedman. The Bayesian structural EM algorithm. In *Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence*, pages 129–138. Morgan Kaufmann Publishers Inc., 1998.
- N. Friedman, D. Geiger, and M. Goldszmidt. Bayesian network classifiers. *Machine Learning*, 29(2-3):131–163, 1997.
- J. Gama and G. Castillo. Learning with local drift detection. In *International Conference on Advanced Data Mining and Applications*, pages 42–55. Springer, 2006.
- J. Gama, I. Žliobaité, A. Bifet, M. Pechenizkiy, and A. Bouchachia. A survey on concept drift adaptation. *ACM Computing Surveys*, 46(4):44, 2014.

- E. S. Gelsema. Abductive reasoning in Bayesian belief networks using a genetic algorithm. *Pattern Recognition Letters*, 16(8):865–871, 1995.
- E. Gibaja and S. Ventura. A tutorial on multi-label learning. *ACM Computing Surveys*, 47, 2015.
- S. Godbole and S. Sarawagi. Discriminative methods for multi-labeled classification. In *Proceedings of the 8th Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 22–30. Springer, 2004.
- D.-J. Guan. Generalized gray codes with applications. *Proceedings of the National Science Council of the Republic of China. Part A, Physical Science and Engineering*, 22(6):841–848, 1998.
- G. W. Hart. Nonintrusive appliance load monitoring. *Proceedings of the IEEE*, 80(12):1870–1891, 1992.
- L. Hawarah, S. Ploix, and M. Jacomino. User behavior prediction in energy consumption in housing using Bayesian networks. In *Proceedings of the 10th International Conference on Artificial Intelligence and Soft Computing*, pages 372–379. Springer, 2010.
- M. Henrion. Propagating uncertainty in Bayesian networks by probabilistic logic sampling. In *Machine Intelligence and Pattern Recognition*, volume 5, pages 149–163. Elsevier, 1988.
- M. Hermann, T. Pentek, and B. Otto. Design principles for industrie 4.0 scenarios. In *Proceedings of the 49th Hawaii International Conference on System Sciences*, pages 3928–3937. IEEE, 2016.
- D. V. Hinkley. Inference about the change-point from cumulative sum tests. *Biometrika*, 58(3):509–523, 1971.
- S. Højsgaard. Graphical independence networks with the gRain package for R. *Journal of Statistical Software*, 46(10):1–26, 2012.
- S. Huang, W. Zuo, and M. D. Sohn. A Bayesian network model for predicting cooling load of commercial buildings. In *Building Simulation*, volume 11, pages 87–101. Springer, 2018.
- E. Hüllermeier, J. Fürnkranz, W. Cheng, and K. Brinker. Label ranking by learning pairwise preferences. *Artificial Intelligence*, 172(16-17):1897–1916, 2008.

- F. Hutter, H. H. Hoos, and T. Stützle. Efficient stochastic local search for MPE solving. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, pages 169–174, 2005.
- J. S. Ide and F. G. Cozman. Random generation of Bayesian networks. In *Brazilian Symposium on Artificial Intelligence*, pages 366–376. Springer, 2002.
- K. Kask and R. Dechter. Mini-bucket heuristics for improved search. In *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence*, pages 314–323. Morgan Kaufmann Publishers Inc., 1999.
- K. Kask and R. Dechter. A general scheme for automatic generation of search heuristics from specification dependencies. *Artificial Intelligence*, 129(1-2):91–131, 2001.
- R. Kohavi. Scaling up the accuracy of naive-Bayes classifiers: A decision-tree hybrid. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*, volume 96, pages 202–207, 1996.
- D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. The MIT Press, 2009.
- X. Kong and S. Y. Philip. An ensemble-based approach to fast classification of multi-label data streams. In *Proceedings of the 7th International Conference on Collaborative Computing: Networking, Applications and Worksharing*, pages 95–104. IEEE, 2011.
- B. Krawczyk, L. L. Minku, J. Gama, J. Stefanowski, and M. Woźniak. Ensemble learning for data stream analysis: A survey. *Information Fusion*, 37:132–156, 2017.
- J. B. Kruskal. On the shortest spanning subtree of a graph and the traveling salesman problem. *Proceedings of the American Mathematical Society*, 7(1):48–50, 1956.
- S. Kullback. *Information Theory and Statistics*. Courier Corporation, 1997.
- J. Kwisthout. Most probable explanations in Bayesian networks: Complexity and tractability. *International Journal of Approximate Reasoning*, 52(9):1452–1469, 2011.
- N. Landwehr, M. Hall, and E. Frank. Logistic model trees. *Machine Learning*, 59(1-2):161–205, 2005.
- P. Langley and S. Sage. Induction of selective Bayesian classifiers. In *Proceedings of the 10th Conference on Uncertainty in Artificial Intelligence*, pages 399–406. Elsevier, 1994.

- P. Langley and S. Sage. Tractable average-case analysis of naïve Bayesian classifiers. In *Proceedings of the 16th International Conference on Machine Learning*, volume 99, pages 220–228, 1999.
- Y. Lee and S. Cho. An efficient energy management system for Android phone using Bayesian networks. In *Proceedings of the 32nd International Conference on Distributed Computing Systems Workshops*, pages 102–107. IEEE, 2012.
- Z. Li and B. D’Ambrosio. An efficient approach for finding the MPE in belief networks. In *Uncertainty in Artificial Intelligence*, pages 342–349. Elsevier, 1993.
- R. Marinescu and R. Dechter. AND/OR branch-and-bound search for combinatorial optimization in graphical models. *Artificial Intelligence*, 173(16-17):1457–1491, 2009.
- E. L. Mencía and J. Fürnkranz. Efficient multilabel classification algorithms for large-scale problems in the legal domain. In *Semantic Processing of Legal Texts*, pages 192–215. Springer, 2010.
- L. L. Minku, A. P. White, and X. Yao. The impact of diversity on online ensemble learning in the presence of concept drift. *IEEE Transactions on Knowledge and Data Engineering*, 22(5):730–742, 2010.
- M. Minsky. Steps toward artificial intelligence. *Proceedings of the Institute of Radio Engineers*, 49(1):8–30, 1961.
- T. A. Nguyen and M. Aiello. Energy intelligent buildings based on user activity: A survey. *Energy and Buildings*, 56:244–257, 2013.
- U. Nodelman, C. R. Shelton, and D. Koller. Continuous time Bayesian networks. In *Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence*, pages 378–387. Morgan Kaufmann Publishers Inc., 2002.
- J. R. Norris. *Markov chains*. Number 2. Cambridge University Press, 1998.
- J. Ortigosa-Hernández, J. D. Rodríguez, L. Alzate, M. Lucania, I. Inza, and J. A. Lozano. Approaching sentiment analysis by using semi-supervised learning of multi-dimensional classifiers. *Neurocomputing*, 92:98–115, 2012.
- N. C. Oza. Online bagging and boosting. In *Proceedings of the 2005 IEEE International Conference on Systems, Man and Cybernetics*, volume 3, pages 2340–2345. IEEE, 2005.
- E. S. Page. Continuous inspection schemes. *Biometrika*, 41(1/2):100–115, 1954.

- S. Park and J. Fürnkranz. Multi-label classification with label constraints. In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases Workshop on Preference Learning*, pages 157–171, 2008.
- A. Pastink and L. C. van der Gaag. Multi-classifiers of small treewidth. In *European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty*, pages 199–209. Springer, 2015.
- J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers, 1988.
- M. Qazi, G. Fung, S. Krishnan, R. Rosales, H. Steck, R. B. Rao, D. Poldermans, and D. Chandrasekaran. Automated heart wall motion abnormality detection from ultrasound images using Bayesian networks. In *International Joint Conference on Artificial Intelligence*, volume 7, pages 519–525, 2007.
- W. Qu, Y. Zhang, J. Zhu, and Q. Qiu. Mining multi-label concept-drifting data streams using dynamic classifier ensemble. In *Proceedings of the 1st Asian Conference on Machine Learning*, pages 308–321. Springer, 2009.
- J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81–106, 1986.
- J. Read, B. Pfahringer, and G. Holmes. Multi-label classification using ensembles of pruned sets. In *Proceedings of the 8th IEEE International Conference on Data Mining*, pages 995–1000. IEEE, 2008.
- J. Read, B. Pfahringer, G. Holmes, and E. Frank. Classifier chains for multi-label classification. *Machine Learning*, 85(3):333–359, 2011.
- J. Read, A. Bifet, G. Holmes, and B. Pfahringer. Scalable and efficient multi-label classification for evolving data streams. *Machine Learning*, 88(1-2):243–272, 2012.
- J. Read, P. Reutemann, B. Pfahringer, and G. Holmes. MEKA: A multi-label/multi-target extension to Weka. *Journal of Machine Learning Research*, 17(21):1–5, 2016.
- J. Rissanen. Modeling by shortest data description. *Automatica*, 14(5):465–471, 1978.
- M. Rüßmann, M. Lorenz, P. Gerbert, M. Waldner, J. Justus, P. Engel, and M. Harnisch. Industry 4.0: The future of productivity and growth in manufacturing industries. 2015. URL [https://www.bcg.com/publications/2015/engineered\\_products\\_project\\_business\\_industry\\_4\\_future\\_productivity\\_growth\\_manufacturing\\_industries.aspx](https://www.bcg.com/publications/2015/engineered_products_project_business_industry_4_future_productivity_growth_manufacturing_industries.aspx).

- R. W. Robinson. Counting labeled acyclic digraphs. In *New Directions in the Theory of Graphs (Ed. F. Harary)*, pages 239–273. Academic Press, New York, 1973.
- J. D. Rodríguez, A. Pérez, D. Arteta, D. Tejedor, and J. A. Lozano. Using multidimensional Bayesian network classifiers to assist the treatment of multiple sclerosis. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(6):1705–1715, 2012.
- J. D. Rodríguez and J. A. Lozano. Multi-objective learning of multi-dimensional Bayesian classifiers. In *Proceedings of the 8th International Conference on Hybrid Intelligent Systems*, pages 501–506. IEEE Computer Society, 2008.
- C. Rojas-Guzman and M. A. Kramer. Galgo: A genetic algorithm decision support tool for complex uncertain systems modeled with Bayesian belief networks. In *Uncertainty in Artificial Intelligence*, pages 368–375. Elsevier, 1993.
- M. Sahami. Learning limited dependence Bayesian classifiers. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*, volume 96, pages 335–338, 1996.
- E. Santos. On the generation of alternative explanations with implications. In *Proceedings of the 7th Conference on Uncertainty in Artificial Intelligence*, page 339. Elsevier, 2014.
- R. E. Schapire and Y. Singer. Improved boosting algorithms using confidence-rated predictions. *Machine Learning*, 37(3):297–336, 1999.
- R. E. Schapire and Y. Singer. BoosTexter: A boosting-based system for text categorization. *Machine Learning*, 39(2-3):135–168, 2000.
- M. Scutari. Learning Bayesian networks with the bnlearn R package. *Journal of Statistical Software*, 35(3):1–22, 2010.
- P. H. Shaikh, N. B. M. Nor, P. Nallagownden, I. Elamvazuthi, and T. Ibrahim. A review on optimized control systems for building energy and comfort management of smart sustainable buildings. *Renewable and Sustainable Energy Reviews*, 34:409–429, 2014.
- Z. Shi, Y. Wen, C. Feng, and H. Zhao. Drift detection for multi-label data streams based on label grouping and entropy. In *Proceedings of the 2014 IEEE International Conference on Data Mining Workshop*, pages 724–731. IEEE, 2014a.

- Z. Shi, Y. Wen, Y. Xue, and G. Cai. Efficient class incremental learning for multi-label classification of evolving data streams. In *Proceedings of the 2014 International Joint Conference on Neural Networks*, pages 2093–2099. IEEE, 2014b.
- S. E. Shimony. Finding MAPs for belief networks is NP-hard. *Artificial Intelligence*, 68(2):399–410, 1994.
- S. E. Shimony and W. Charniak. A new algorithm for finding MAP assignments to belief networks. In *Proceedings of the 6th Annual Conference on Uncertainty in Artificial Intelligence*, pages 185–196. Elsevier Science Inc., 1990.
- T. Shoji, W. Hirohashi, Y. Fujimoto, Y. Amano, S. Tanabe, and Y. Hayashi. Personalized energy management systems for home appliances based on Bayesian networks. *Journal of International Council on Electrical Engineering*, 5(1):64–69, 2015.
- G. Song and Y. Ye. A new ensemble method for multi-label data stream classification in non-stationary environment. In *Proceedings of the 2014 International Joint Conference on Neural Networks*, pages 1776–1783. IEEE, 2014.
- F. Stella and Y. Amer. Continuous time Bayesian network classifiers. *Journal of Biomedical Informatics*, 45(6):1108–1119, 2012.
- B. K. Sy. Reasoning MPE to multiply connected belief networks using message passing. In *Proceedings of the 10th National Conference on Artificial Intelligence*, pages 570–576, 1992.
- M. Teyssier and D. Koller. Ordering-based search: A simple and effective algorithm for learning Bayesian networks. *Proceedings of the 21st Conference on Uncertainty in Artificial Intelligence*, pages 584–590, 2005.
- P. Trajdos and M. Kurzynski. Multi-label stream classification using extended binary relevance model. In *Proceedings of the 2015 IEEE Trustcom/BigDataSE/ISPA*, pages 205–210. IEEE Computer Society, 2015.
- A. Tsanas and A. Xifara. Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools. *Energy and Buildings*, 49:560–567, 2012.
- G. Tsoumakas, I. Katakis, and I. Vlahavas. Mining multi-label data. In *Data Mining and Knowledge Discovery Handbook*, pages 667–685. Springer, 2009.

- L. C. van der Gaag and P. R. de Waal. Multi-dimensional Bayesian network classifiers. In *Proceedings of the 3rd European Workshop in Probabilistic Graphical Models*, pages 107–114, 2006.
- L. Wang, H. Shen, and H. Tian. Weighted ensemble classification of multi-label data streams. In *Proceedings of the 21st Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 551–562. Springer, 2017.
- P. Wang, P. Zhang, and L. Guo. Mining multi-label data streams using ensemble-based active learning. In *Proceedings of the 12th SIAM International Conference on Data Mining*, pages 1131–1140. SIAM, 2012.
- Q. Wei, Z. Yang, Z. Junping, and W. Yong. Mining multi-label concept-drifting streams using ensemble classifiers. In *Proceedings of the 6th International Conference on Fuzzy Systems and Knowledge Discovery*, volume 5, pages 275–279. IEEE, 2009.
- G. Widmer and M. Kubat. Learning in the presence of concept drift and hidden contexts. *Machine Learning*, 23(1):69–101, 1996.
- D. H. Wolpert and W. G. Macready. No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1(1):67–82, 1997.
- E. S. Xioufis, M. Spiliopoulou, G. Tsoumakas, and I. P. Vlahavas. Dealing with concept drift and class imbalance in multi-label stream classification. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence*, pages 1583–1588, 2011.
- Y. Yang. An evaluation of statistical approaches to text categorization. *Information Retrieval*, 1(1-2):69–90, 1999.
- Y. Yang and X. Liu. A re-examination of text categorization methods. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 42–49. ACM, 1999.
- J. H. Zaragoza, L. E. Sucar, and E. F. Morales. A two-step method to learn multidimensional Bayesian network classifiers based on mutual information measures. In *24th International FLAIRS Conference*, pages 644–649, 2011.
- M. Zhang and Z. Zhou. ML-KNN: A lazy learning approach to multi-label learning. *Pattern Recognition*, 40(7):2038–2048, 2007.
- M. Zhang and Z. Zhou. A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 26(8):1819–1837, 2014.

- X. Zhang, Q. Yuan, S. Zhao, W. Fan, W. Zheng, and Z. Wang. Multi-label classification without the multi-label cost. In *Proceedings of the 10th SIAM International Conference on Data Mining*, pages 778–789. SIAM, 2010.
- H. Zhao and F. Magoulès. A review on the prediction of building energy consumption. *Renewable and Sustainable Energy Reviews*, 16(6):3586–3592, 2012.
- Z. Zheng and G. I. Webb. Lazy learning of Bayesian rules. *Machine Learning*, 41(1):53–84, 2000.
- M. Zhu, S. Liu, and J. Jiang. A hybrid method for learning multi-dimensional Bayesian network classifiers based on an optimization model. *Applied Intelligence*, 44(1):123–148, 2016.
- S. Zhu, X. Ji, W. Xu, and Y. Gong. Multi-labelled classification using maximum entropy method. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 274–281. ACM, 2005.

# Glosario

**Cambio de concepto** En un contexto de aprendizaje supervisado de flujos de datos, el cambio a lo largo del tiempo de la relación entre las variables predictoras y la(s) variable(s) clase se denota como cambio de concepto.

**CB-MBC** Un MBC clase-puente descomponible, en inglés *class-bridge decomposable* MBC, es un MBC de topología restringida sobre el que un problema de inferencia se puede descomponer en sub-problemas más fáciles.

**DAG** Del inglés, *directed acyclic graph*, es un grafo dirigido que no tiene ciclos. La estructura de una red Bayesiana es un DAG.

**Ensemble** Un *ensemble* es un conjunto de clasificadores cuyas predicciones se combinan para predecir nuevas instancias. Es una técnica muy popular como meta-clasificador y también en un contexto de clasificación de flujos de datos para hacer frente al potencial cambio de concepto.

**Flujo de datos** Un escenario de clasificación de flujos de datos se diferencia de un escenario tradicional estacionario por su carácter dinámico, i.e., se reciben nuevos datos continuamente, lo que induce posibles cambio de concepto a lo largo del tiempo y requerimientos especiales de tiempo y de memoria.

**Industria 4.0** Cuarta revolución industrial que consiste en la introducción de las tecnologías digitales en la industria.

**MAP** El **máximo a posteriori** es un proceso de inferencia abductiva que busca las configuraciones más probables, i.e., que mejor explican la evidencia.

**MBC** Un clasificador multi-dimensional de redes Bayesianas, en inglés *multi-dimensional Bayesian network classifier*, es una red Bayesiana especialmente diseñada para resolver problemas de clasificación multi-dimensional.

**MBCTree** Clasificador multi-dimensional híbrido propuesto en este trabajo que consiste en un árbol de clasificación con MBCs en los nodos hoja.

**MPE** La explicación más probable, en inglés *most probable explanation*, también llamada abducción total, es un tipo de inferencia MAP tal que se busca la configuración más probable de todas las variables no observadas, i.e., aquellas que no forman parte de la evidencia.

**Multi-label** Un problema de clasificación *multi-label* o multi-etiqueta consiste en predecir qué etiquetas corresponden a una nueva instancia. Se trata de una configuración específica del problema más general de clasificación multi-dimensional, tal que cada variable clase en relación a una etiqueta toma únicamente dos valores en base a la correspondencia o no de la etiqueta.

**Wrapper** Un método de aprendizaje *wrapper* es dependiente del propio algoritmo de aprendizaje a diferencia de los métodos *filter*, tal que se busca mejorar una medida de bondad del modelo en construcción, generalmente el *accuracy*. Los métodos *wrapper* suelen obtener mejores resultados a costa de un mayor coste computacional.

## **Anexos**



# Anexo A

## Gestión del proyecto

### A.1. Planificación

En el lanzamiento del proyecto se detallaron el alcance y los objetivos del mismo (ver Sección 1.5). De igual manera, se realizó una planificación temporal de las principales actividades que conforman los objetivos marcados (Figura A.1). La fecha de finalización del proyecto se marcó el día 17 de julio de 2018 en base a los plazos disponibles para el depósito del Trabajo Fin de Máster.

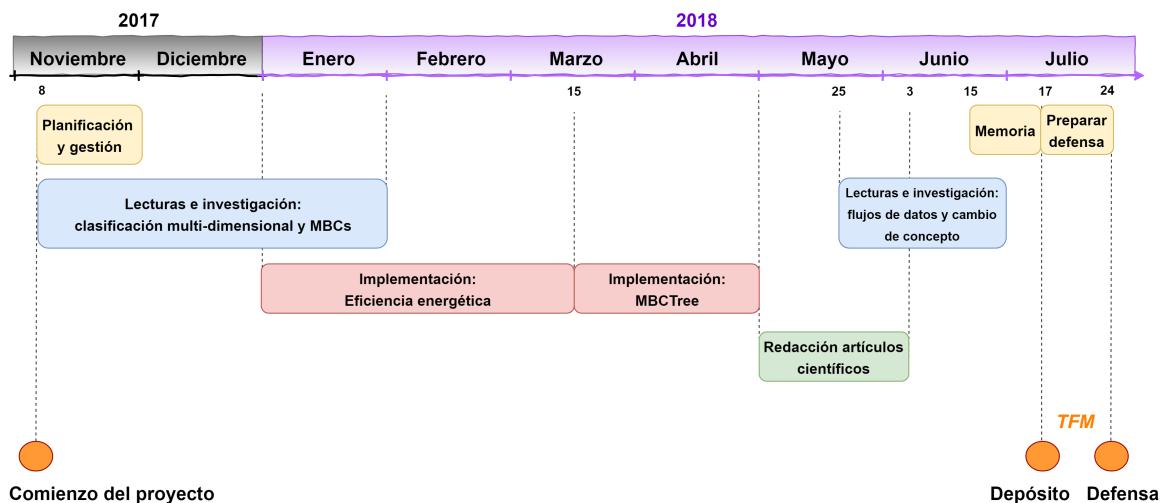


Figura A.1: Planificación temporal del proyecto.

### A.2. Metodología

El proyecto realizado ha estado enfocado en un ámbito de investigación. Ha sido necesaria la lectura de toda la literatura relevante relacionada con el proyecto.

Para evaluar de forma continua el progreso del trabajo se celebraba una reunión de treinta minutos todos los viernes cada dos semanas con los tutores del proyecto Pedro Larrañaga y Concha Bielza. En estas reuniones se abordaba principalmente la

evaluación y discusión de los resultados obtenidos desde la anterior reunión, así como la proyección y planteamiento de nuevos pasos y objetivos.

Con una frecuencia mucho menor, sumando un total de tres reuniones a lo largo del transcurso del trabajo, los directores de innovación Patxi Samaniego y Javier Díaz de la empresa Etxe-Tar acudían a nuestra Escuela para discutir los resultados obtenidos en relación al problema de eficiencia energética. Como expertos del dominio, juzgaban estos resultados y extraían conclusiones, además de plantear nuevos caminos para explorar en base a lo que se concluía anteriormente.

### A.3. Herramientas utilizadas

Durante la realización de este trabajo se han utilizado las siguientes herramientas:

- *Python* y *R*: lenguajes de programación base del proyecto, tanto para la implementación y evaluación del modelo propuesto MBCTree como para las aplicaciones a la eficiencia energética en la Industria 4.0.
- *bnlearn* [[Scutari, 2010](#)] y *gRain* [[Højsgaard, 2012](#)]: paquetes de R para el aprendizaje e inferencia en redes Bayesianas.
- *Weka* [[Frank et al., 2006](#)] y *Meka* [[Read et al., 2016](#)]: plataformas de software para el aprendizaje automático y la minería de datos, permitiendo resolver problemas multi-dimensionales la segunda.
- *Git* y *GitHub*: software y plataforma de desarrollo para el control de versiones.
- *Google Scholar*: motor de búsqueda para la obtención de toda la literatura académica utilizada como apoyo a la investigación del proyecto.
- *LaTeX*: sistema de composición de textos para la documentación del proyecto.
- *Beamer*: clase de LaTeX para crear la presentación del proyecto.

